Automated Development of Clinical Prediction Models Using Genetic Programming

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Christian A. Bannister

September 2015

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed . . . . . (candidate) Date . . . . . .

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed . . . . . (candidate) Date . . . . . .

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed . . . . . (candidate) Date . . . . . .

Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed . . . . . (candidate) Date . . . . . .

Statement 4: Previously Approved Bar On Access

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed . . . . . (candidate) Date . . . . . .

To Marie, Teilo & Ellie
for their patience and support.

Summary

Genetic programming is an Evolutionary Computing technique, inspired by biological evolution, capable of discovering complex non-linear patterns in large datasets. Despite the potential advantages of genetic programming over standard statistical methods, its applications to survival analysis are at best rare, primarily because of the difficulty in handling censored data. The aim of this work was to develop a genetic programming approach for survival analysis and demonstrate its utility for the automatic development of clinical prediction models using cardiovascular disease as a case study.

We developed a tree-based untyped steady-state genetic programming approach for censored longitudinal data, comparing its performance to the de facto statistical method —Cox regression—in the development of clinical prediction models of future cardiovascular events in patients with symptomatic and asymptomatic cardiovascular disease, using large observational datasets.

Results showed that Cox regression and the developed genetic programming approach produced similar results when evaluated in common validation datasets. Despite generally comparable performance, albeit in slight favour of the Cox model, the predictors selected for representing their relationships with the outcome were quite different and, on average, the models developed using genetic programming used considerably fewer predictors. The genetic programming models were more complex and thus more difficult to validate by domain experts, however these models were developed in an automated fashion, using fewer input variables, without the need for domain specific knowledge and expertise required to appropriately perform survival analysis.

This work has demonstrated the strong potential of genetic programming as a methodology for automated development of clinical prediction models for diagnostic and prognostic research, where the primary goal is accurate prediction. In aetiological research, where the primary goal is to examine the relative strength of association between risk factors and the outcome, then Cox regression and its variants remain as the de facto approach.

Abstract

Background & Aims Genetic programming is an Evolutionary Computing technique, inspired by biological evolution, capable of discovering complex non-linear patterns in large datasets. Genetic programming is a general methodology, the specific implementation of which requires development of several different specific elements such as problem representation, fitness, selection and genetic variation. Despite the potential advantages of genetic programming over standard statistical methods, its applications to survival analysis are at best rare, primarily because of the difficulty in handling censored data. The aim of this work was to develop a genetic programming approach for survival analysis and demonstrate its utility for the automatic development of clinical prediction models using cardiovascular disease as a case study.

Methods We developed a tree-based untyped steady-state genetic programming approach for censored longitudinal data, comparing its performance to the *de facto* statistical method—Cox regression—in the development of clinical prediction models for the prediction of future cardiovascular events in patients with symptomatic and asymptomatic cardiovascular disease, using large observational datasets. We also used genetic programming to examine the prognostic significance of different risk factors together with their non-linear combinations for the prognosis of health outcomes in cardiovascular disease.

Results These experiments showed that Cox regression and the developed steady-state genetic programming approach produced similar results when evaluated in common validation datasets. Despite slight relative differences, both approaches demonstrated an acceptable level of discriminative and calibration at a range of times points. Whilst the application of genetic programming did not provide more accurate representations of factors that predict the risk of both symptomatic and asymptomatic cardiovascular disease when compared with existing meth-

ods, genetic programming did offer comparable performance. Despite generally comparable performance, albeit in slight favour of the Cox model, the predictors selected for representing their relationships with the outcome were quite different and, on average, the models developed using genetic programming used considerably fewer predictors. The results of the genetic programming confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling; age, previous atherosclerosis, and albumin for secondary prevention; age, recorded diagnosis of 'other' cardiovascular disease, and ethnicity for primary prevention in patients with type 2 diabetes. When considered as a whole, genetic programming did not produce better performing clinical prediction models, rather it utilised fewer predictors, most of which were the predictors that Cox regression estimated be most strongly associated with the outcome, whilst achieving comparable performance. This suggests that genetic programming may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

Conclusions To our knowledge, this work is the first study to develop a genetic programming approach for censored longitudinal data and assess its value for clinical prediction in comparison with the well-known and widely applied Cox regression technique. Using empirical data this work has demonstrated that clinical prediction models developed by steady-state genetic programming have predictive ability comparable to those developed using Cox regression. The genetic programming models were more complex and thus more difficult to validate by domain experts, however these models were developed in an automated fashion, using fewer input variables, without the need for domain specific knowledge and expertise required to appropriately perform survival analysis. This work has demonstrated the strong potential of genetic programming as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes in the presence of censored data. This work compared untuned genetic programming models that were developed in an automated fashion with highly tuned Cox regression models that was developed in a very involved manner that required a certain amount of clinical and statistical expertise. Whilst the highly tuned Cox regression models performed slightly better in validation data, the performance of the automatically generated genetic programming models were generally comparable. The comparable performance demonstrates the utility of genetic programming for clinical prediction modelling and prognostic research, where the primary goal is accurate prediction. In aetiological research, where the primary goal is to

examine the relative strength of association between risk factors and the outcome, then Cox regression and its variants remain as the *de facto* approach.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Irena Spasić, for her continued support of my PhD research, for her patience, guidance, and encouragement. Her advice on my research as well as my career have been greatly appreciated. Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Craig Currie, Prof. Glyn Elwyn, and Prof. Alun Preece, for their insightful comments and encouragement, but also for posing the difficult questions that have ultimately broadened and improved my research. I would especially like to thank Prof. Craig Currie and his team for the opportunity to join them, for making me feel so welcome, and for sharing with me some of their considerable knowledge and experience. Without their precious support it would not be possible to conduct this research. I would like to thank Dr. Chris Poole, not only for his original ideas and enthusiasm, but for his much valued mentorship. I would also like to thank Prof. Joshua Knowles and Dr. Jianhua Shao for serving as my review panel and for letting defence of my research be an enjoyable moment. This dissertation would not have been possible without funding from the Cardiff School of Computer Science & Informatics and the Medical Research Council.

A special thank you to my family. Words cannot express how grateful I am to my parents, my partner Marie, and my two wonderful children, Teilo and Ellie for all the sacrifices that they have made on my behalf and for a constant source of motivation.

Contents

List of Publications

The work introduced in this thesis is based on the following publications.

- Bannister CA, Poole CD, Jenkins-Jones S, Morgan CLl, Elwyn G, Spasić I, Currie CJ. External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations. Diabetes Care 2014;37:537-45. http://dx.doi.org/10.2337/dc13-1159

- Bannister CA, Poole CD, Jenkins-Jones S, Morgan CLl, Elwyn G, Currie CJ. Validation of UKPDS Risk Engine Predictions Among Patients with Type 2 Diabetes Routinely Managed in UK Primary Care. Diabetes 2013;62(suppl 1):A1-A98 276-OR. http://dx.doi.org/10.2337/db13-1-387

- Bannister CA, Currie CJ, Preece A, Spasić I. Automatic development of clinical prediction models with genetic programming: a case study in cardiovascular disease. Value Health 2014;17:A200-1. http://dx.doi.org/10.1016/j.jval.2014.03.1171

List of Figures

List of Tables

List of Algorithms

List of Acronyms

AAA Abdominal Aortic Aneurysm

ACE Angiotensin-Converting-Enzyme

AFPO Age-Fitness Pareto Optimisation

AFT Accelerated Failure Time

AI Artificial Intelligence

ANN Atrifical Neural Network

APT Anti-Platelet Therapy

ARB Angiotensin Receptor Blockers

AUC area under receiver operating characteristic curve

BMI Body Mass Index

BNF British National Formulary

CD Crowding Distance

CHD Coronary Heart Disease

CPRD Clinical Practice Research Datalink

CVD Cardiovascular Disease

DBP Dyastolic Blood Pressure

DoH Department for Health

EA Evolutionary Algorithms

EC Evolutionary Computation

EMOA Evolutionary Multi-Objective Algorithms

EP Evolutionary Programming

EPR Electronic Patient Records

ES Evolutionary Strategies

FTIR Fourier Transform Infrared Spectroscopy

GA Genetic Algorithms

GCMS Gas Chromatography-Mass Spectrometry

GMOGP Generational Multi-Objective Genetic Programming

GP Genetic Programming

GSOGP Generational Single-Objective Genetic Programming

HDL High-density Lipoprotein Cholesterol

HES Hospital Episode Statistics

ICD International Classification of Diseases

IMT Intima Media Thickness

ISAC Independent Scientific Advisory Committee

IQR Inter-Quartile Range

KM Kaplan-Meier

LDL Low-density Lipoprotein Cholesterol

LL Log Likelihood

MHRA Medicines and Healthcare products Regulatory Agency

MI Myocardial Infarction

MICE Multivariate Imputation by Chained Equations

ML Machine Learning

MLE Maximum Likelihood Estimation

MOGP Multi-Objective Genetic Programming

NDS Non-Dominated Sorting

NHS National Health Service

NICE National Institute of Clinical Excellence

ONS Office of National Statisitcs

OHA Oral Hypoglycaemic Agent

OPCS Office of Population Censuses and Surveys, Classification of Surgical Operations and Procedure

PE Pulmonary Embolism

PCT Primary Care Trusts

PH Proportional Hazards

QOF Quality and Outcomes Framework

RCT Randomised Clinical Trials

SBP Systolic Blood Pressure

SES Socio-Economic Status

SHA Strategic Health Authorities

SMART Second Manifestations of ARTerial disease

SNP Single Nuclear Polymorphisms

SMOGP Steady-state Multi-Objective Genetic Programming

SSOGP Steady-state Single-Objective Genetic Programming

T2DM Type II Diabetes Mellitus

TC Total Cholesterol

THIN The Health Improvement Network

UKPDS United Kingdom Prospective Diabetes Study

UKPDS-RE UKPDS Risk Engine

UTS Up-To-Standard

WHO World Health Organisation

Chapter 1

Introduction

Prognosis is central to medicine. All diagnostic and therapeutic actions aim to improve prognosis [286]. Physicians and health policy makers need to make predictions on the prognosis of a disease in their decision making. Clinical prediction models provide inputs for this decision making by providing estimates of individual probabilities or risks and benefits [182]. Clinical prediction models combine a variety of characteristics to predict some diagnostic or prognostic outcome [286]. Current methods for clinical prediction use traditional statistical techniques, often using relatively small samples. Generally statistical approaches have inherent restrictions on the complexity of patterns they can learn and the volume of data they can handle whilst remaining effective.

Generally, survival analysis is a collection of statistical procedures for the analysis of data in which the outcome of interest (the survival outcome) is the time to event, which is typically referred to as survival time [162]. More specifically, survival analysis involves the estimation of the distribution of the time it takes for an event to happen to a patient based on some set of features, which are also known as explanatory variables, predictors or covariates. The event may for example be death, disease incidence, or recurrence. A key characteristic of survival data is that the follow-up of patients is typically incomplete [286]. For example, some patients may have been followed for 5 days, some for 15 days, etc, and we may be interested in predicting 30-day survival. Such incomplete data, is what we call *censored data*. In essence, censoring occurs when we have some information about an individual survival times, but we do not know the survival time exactly in all subjects. Survival is an important long-term outcome in prognostic research, including medical research areas such as cardiology and oncology.

1.1 Motivation

Traditionally the Kaplan-Meier (KM) method, a non-parametric approach, has been used for exploratory analysis of survival data. Using the KM method survival curves can be generated for various subgroups (e.g. females versus males) to investigate the effect of explanatory variables on survival. However, the KM approach is limited as it is unable to consider the effect of multiple explanatory variables simultaneously. To overcome this limitation, several regression modelling approaches have been proposed to enable prediction of time to event in the presence of censored data [287]. Most of these models have come from the long-established statistical literature such as parametric survival models, including the Weibull, lognormal and Gompertz models and the semi-parametric Proportional Hazards model proposed by Cox [48]. Hereafter, these will be jointly referred to as linear statistical methods or models. In medical and epidemiological studies the Cox Proportional Hazards model (or Cox regression) is the most often used model for survival outcomes.

Alternate methods for survival analysis may be based on machine learning, e.g. Atrifical Neural Network (ANN). These will be referred to as non-linear statistical methods or models. There have been several studies that have compared such novel non-linear statistical methods with their classic linear counterparts for survival outcomes [148, 138, 229, 248]. However the results are mixed as to whether these non-linear methods offer improved performance. For example Schwarzer et al. [269] reviewed a substantial number of studies which have used ANNs in the diagnostic and prognostic classification in cancer, concluding that there is no evidence so far that application of ANNs represents real progress in the field of diagnosis and prognosis in oncology. Sargent [269] has also reviewed a number of these comparison studies showing that the majority have claimed equal performance but could not rule out the possibility of bias. GP, however, is a relatively new non-linear method that may improve the selection and transformation of predictors, and it may lead to models with good predictive accuracy in new patients [22, 89, 136, 238, 297].

GP is an Evolutionary Computation (EC) technique inspired by population genetics, and evolution at the population level, as well as the Mendelian understanding of the structure and mechanisms [29, 195, 278]. GP automatically solves complex problems without requiring the user

to know or specify the form or structure of the solution in advance [239]. This makes GP well suited to symbolic regression, where in addition to searching for the solution to the complex associations between predictors and outcome, GP also searches for the optimal model structure. This in turn makes GP well suited to prediction, primarily an estimation problem, where the mutual correlations between predictors and the outcome are to be estimated. GP has been shown to work well for recognition of structures in large data sets and has the intrinsic advantage of automatically selecting a subset of inputs or features during the evolutionary process [212, 306].

Because GP has no fixed model structure it can represent complex non-linear associations that could not be achieved using linear regression techniques, and theoretically achieve higher predictive accuracy. However, the flexibility of regression models can be greatly enhanced through the use of factional polynomial, restricted cubic splines and interaction terms, potentially increasing its predictive accuracy [22, 111, 112, 164]. Although it has be emphasised that this is not normally done, as it complicates interpretation and the correct use of the appropriate regression methods is quite difficult, as they require extensive statistical knowledge [271].

Despite its potential, critics state that GP is more prone to over-fitting compared to conventional development methods [136]. An often cited argument against the use of of machine learning techniques in clinical research, is that modelled relationship between predictors and outcome can be highly complex and thus difficult to validate by domain experts, sometimes referred to 'black-box' techniques. However, unlike many machine learning approaches, genetic programming produces an explicit human-understandable model and is not a 'black box' method. GP has been used in medical research for classification and, to a lesser extend, prediction. However its value for prediction on censored data, for survival analysis, has not yet been documented.

1.2 Goals of this Work

The main hypothesis of this research is that the application of GP can provide more accurate representation of factors that predict the risk of Cardiovascular Disease (CVD) when compared with existing methods. That is, the development and validation of a GP approach for survival

analysis would offer improved performance when compared to the *de facto* statistical methods for clinical prediction modelling in censored longitudinal data.

There are three main goals of this work. Firstly, to motivate the need for improved clinical risk prediction methods and models by validating the performance the *de facto* statistical methods in a contemporary real-world clinical setting. Secondly, to demonstrate the utility of GP for the automatic development of clinical prediction models using CVD as case studies. Thirdly, to apply GP to examine the prognostic significance of different risk factors together with their non-linear combinations in order to provide more accurate prognosis of health outcomes in CVD.

1.3 Contributions of this Work

This thesis makes six main contributions:

1. *The de facto cardiovascular risk prediction models for T2DM may be unsuitable* Using data from Clinical Practice Research Datalink (CPRD) this work has preformed the largest, independent, external validation of the *de facto* cardiovascular risk model for people with Type II Diabetes Mellitus (T2DM), the UKPDS Risk Engine (UKPDS-RE), in a diverse and contemporary setting. This work showed poor performance, suggesting that the UKPDS-RE is not suitable for predicting cardiovascular risk in UK subjects with T2DM. Considering the widespread application of these prediction models, this work suggests a need for revised risk equations in T2DM.

2. *Developement of a GP approach for survial analysis of censored data* GP is a general methodology, the specific implementation of which requires development of several different specific elements such as problem representation, fitness, selection and genetic variation. This work has developed a tree-based untyped Steady-state Single-Objective Genetic Programming (SSOGP) approach for the automated development of clinical prediction models in the presence of censored longitudinal data. Specific GP elements were developed and implemented, such as fitness functions and search heuristics, to handle the problem-specific complexities of censored data and facilitate survival analysis.

3. *Generational GP approaches are too computionally expesive for large observational cohorts* This work attempted to implement and evaluate the utility of two broad classes of GP, steady-state GP common in modern GP systems and the more traditional generational GP approach. Despite considerable effort, when the developed generational approaches were applied to the large observational datasets of censored longitudinal data identified for this work, they failed as a result of requiring more memory that was available in the computing resources allocated for this work. This serves to demonstrate the utility of the relatively computationally efficient steady-state GP approach for analysing large observational cohorts of patients.

4. *GP has utility for the automatic development of clinical prediction models in censored data* Using data from the Second Manifestations of ARTerial disease (SMART) study and from CPRD we have demonstrated that symbolic regression models generated by the developed SSOGP approach had predictive ability comparable to that of the *de facto* statistical method—Cox regression—for the prediction of future cardiovascular events in patients with symptomatic and asymptomatic CVD. These experiments compared untuned SSOGP symbolic regression models that were developed in an automated fashion using only basic parameters settings recommended from the GP literature, with highly tuned Cox regression models that were developed in a very involved manner that required a certain amount of clinical and statistical expertise. Whilst the highly tuned Cox regression models performed slightly better in their validation datasets, the performance of the automatically generated symbolic regression models were generally comparable, and on average consisting of considerably fewer predictors. Using symptomatic and asymptomatic CVD as case studies—for secondary and primary prevention clinical settings, respectively—these findings demonstrate the utility of GP as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes in the presence of censored longitudinal data.

5. *Confirmation of the prognostic significance of certain risk factors in symtomatic CVD* This work has applied GP to examine the prognostic significance of different risk factors together with their non-linear combinations in predicting cardiovascular outcomes in patients with symptomatic and asymptomatic CVD. Whilst the application of GP did

not provide more accurate representations of factors that predict the risk of both symptomatic and asymptomatic CVD when compared with existing methods, GP did offer comparable performance. Despite generally comparable performance, albeit in slight favour of the Cox model, the predictors selected for representing their relationships with the outcome were quite different and, on average, the models developed using GP used considerably fewer predictors. The results of the GP confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling; age, previous atherosclerosis, and albumin for secondary prevention; age, recorded diagnosis of 'other' CVD, and ethnicity for primary prevention in patients with T2DM. When considered as a whole, GP did not produce a better performing clinical prediction model, rather it utilised fewer predictors, most of which were the predictors that the Cox regression estimated be most strongly associated with the outcome, whilst achieving comparable performance. This suggests that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

6. *In practice GP is robust* By implementing SSOGP without model tuning, using only basic parameters values recommended from general GP literature, observing that it has performance comparable to the *de facto* statistical method, we have confirmed the observations of other authors, that in practice GP is robust and likely to work well over a wide range of parameter values.

1.4 Overview of this Thesis

The remainder of this work is structured as follows. Chapter 2 describes the wider context clinical prediction modelling and the UK health system, defining the challenges of predicting risk in the presence of censored data, and provides motivation for the application of GP for cardiovascular risk prediction. Chapter 3 surveys and critically assesses the existing research related to this work. Chapter 4 gives an overview of the essential common themes in the diverse field of GP and discusses the specific methodological elements that form the developed SSOGP approach for censored longitudinal data, which are implemented and assessed in the subsequent experiment chapters. Chapter 5 describes a set of experiments that independently and externally

validated the performance of the *de facto* cardiovascular risk prediction model for patients with T2DM, the UKPDS-RE, using data from CPRD. Chapter 6 describes a set of experiments that demonstrate the utility of the developed SSOGP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events in patients with symptomatic CVD using censored survival data from the SMART study. Chapter 7 describes a set of experiments with a very similar experimental set-up those in the previous chapter, that demonstrate the utility of the developed SSOGP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events, but uses a much larger observational cohort of patients from CPRD in a primary prevention clinical setting, where patients have asymptomatic CVD. Finally in chapter 8, the contributions of this work are revisited, results are discussed, limitations critically assessed, and opportunities for further research identified.

1.5 Overview of Related Publications

Here we give an overview of the way in which parts of this thesis have been published.

Chapter 5: External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations

The content of this chapter is based on research published as an original research article in the Diabetes Care journal [15] and also published in the proceedings of the 73rd Scientific Sessions of the American Diabetes Association (ADA 2013) [16].

Chapter 6: Automatic development of clinical prediction models with genetic programming: a case study in the SMART cohort

A large portion of this chapter is published in the proceedings of the 19th Annual International Meeting of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR 2014) [14] and at the time of writing are under review as an original research article in the Journal of Biomedical Informatics.

Chapter 2

Backgound

We first focus our attention on the wider context clinical prediction modelling and the UK health system, defining the challenges of predicting risk in the presence of censored data, and providing motivation for the application of GP for cardiovascular risk prediction.

First we introduce clinical prediction modelling, the prediction cardiovascular risk in patients with diabetes, and the national health system in the UK. Before introducing the data source selected for the experiments and its associated clinical coding, we give and overview of study design and potential sources of data. We then provide an overview of classic statistical approaches to clinical prediction modelling in the presence of censored data. Before introducing GP, we give an overview of machine learning in the context of EC. Finally we outline the of motivation of this thesis.

2.1 Clinical Predciton Modelling

Prognosis is central to medicine. All diagnostic and therapeutic actions aim to improve prognosis [286]. Physicians and health policy makers need to make predictions on the prognosis of a disease (of the likelihood of an underlying disease) in their decision making. Traditionally, predictions were more implicit and medicine more subjective. However we are now in an era of 'evidence-based medicine' which is defined as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" [232, 261]. Evidence based medicine applies scientific method to clinical practice [102]. Another more recent development is the tendency towards 'shared decision-making' where physicians and

patients both actively participate in deciding on choices for diagnostic tests and therapeutic interventions [47, 286].

Clinical prediction models may provide the evidence-based input for shared decision making, by providing estimates of individual probabilities or risks and benefits [182]. Clinical prediction models have many names including clinical prediction rules, prognostic models, predictive risk models, risk scores, risk equations or nanograms [262]. Clinical prediction models combine a number of characteristics (e.g. features related to the patient, the disease, or treatment) to predict some diagnostic or prognostic outcome [286]. With increasing availability of electronic patient records the interest in prognostic research will further increase because electronic records facilitate the application of predictions rules in clinical practice [22].

The aims of clinical prediction modelling fall into two broad categories, obtaining accurate predictions and obtaining insights into disease mechanisms and pathophysiological processes. From a modelling perspective this is either accurately predicting the probability or risk of the outcome or understanding and quantifying the effect of of the risk factors (features) on the outcome. In this thesis we concern ourselves with the former, obtaining accurate clinical predictions.

2.2 Cardiovacluar Disease & Diabetes

CVD is a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, pulmonary embolism, hypertension and heart failure. The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol [311]. CVD is the leading cause of death globally: [311]. An estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to cerebrovascular disease. By 2030, almost 23.6 million people will die from CVD, mainly from heart disease and cerebrovascular disease. These are projected to remain the single leading causes of death [311].

Diabetes is on the rise, in the UK and around the world [142, 214, 13, 31, 38, 263, 313]. Forecasting models have shown that the prevalence of diabetes is steadily increasing, and that diabetes is not a localised chronic condition [159, 312, 199, 26, 228]. Major contributors to this increase in prevalence are obesity and an ageing population, both of which increase risk of T2DM. T2DM is a condition that affects 8% of the US population [33] and 4% of the UK population [70]. Good glucose control is important to reduce the risk of developing microvascular complications. This is initially achieved through diet and exercise, but glucose-lowering medication is required in most patients with progressing diabetes.

Asymptomatic patients that are suspected to be at high risk need to be identified by General Practitioners so they can offer advice about lifestyle changes and initiate preventative treatment. To facilitate this, General Practitioners need tools that can accurately and reliably predict cardiovascular risk in their patients.

2.3 Cardiovascular Risk

In public health, prediction models may help target preventative interventions to subjects with relatively high risk of having or developing a disease [286]. Numerous models have been developed to predict the future occurrence of disease in asymptomatic subjects in the general population and in specific sub-populations. Arguably, the domain that has seen the most research into the application of prediction models for primary and secondary prevention is CVD.

National policies for the management of both CVD and T2DM advocate the calculation of CVD risk in order to identify high-risk patients for targeted interventions [264, 295, 245, 66, 216]. Several multivariable risk prediction models (or risk scores) have been developed for the general, non-diabetic population that also account for diabetes (see chapter 3), but only a few are specific to T2DM [303]. A minority of of these risk scores have been validated and tested for their predictive accuracy, with only a few showing a discriminative value of $\geq$0.80 [303]. The impact of applying these risk scores in clinical practice is almost completely unknown, but their use is recommended in various national guidelines. Not only are these risks scores advocated for communicating cardiovascular risk to diabetic patients, they are relied upon for public health decisions. Evidence that these equations are inadequate could bring into

question the evidence base underpinning many clinical decisions and public policies about the management of T2DM.

2.4 National Health Systems: UK perspective

Founded in 1948 the National Health Service (NHS) is the shared name for three of the four publicly funded healthcare systems in the United Kingdom. The individual systems are NHS (England), HSENI (Health and Social Care in Northern Ireland), NHS Scotland and NHS Wales. They provide a comprehensive range of health services, the vast majority of which are free at the point of use to residents of the United Kingdom. Throughout this document the 'NHS' or the 'National Health Service' shall refer to the health systems of England, Scotland and Wales, 'NHS (England)' shall refer to the health system in England only. The NHS is the world's largest publicly funded health service, employing over 1.7 million staff [219], with a 2011/2012 budget of £106 billion [67].

Appointed by the Prime Minister, the Secretary of State for Health is responsible for the Department for Health (DoH), which in turn is responsible for the NHS. Strategic Health Authorities (SHA) enact DoH policy at a regional level, enabling directives and implementing fiscal policy. SHAs are also responsible for strategic supervision of the NHS trusts such as hospitals, ambulance services, care trusts, mental health services and primary care trusts. Primary Care Trusts (PCT) provide and/or commission NHS services and form the local management of the NHS, each with their own budgets and priorities. PCTs are responsible for $\approx$80% of the NHS budget funding General Practitioners and prescriptions [108]. However from April 2013, SHAs and PCTs will be abolished, being, replaced by the NHS Commissioning Board and a countrywide network of Clinical Commissioning Groups.

Primary care, in contrast to secondary care, refers to local 'frontline' services acting as a first point of contact (figure 2.1). A range of independent contractors including General Practitioners, walk-in centres, dentists, pharmacists and optometrists deliver the NHS' primary care. Secondary care refers to acute healthcare, emergencies and elective care, which are usually provided in NHS hospitals.

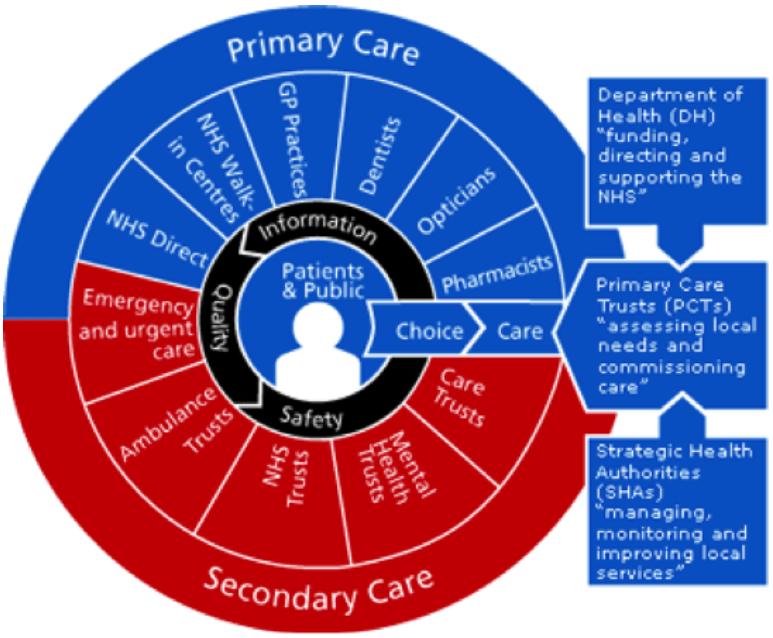


Figure 2.1: Figure illustrating primary and secondary care in the NHS. (`http://www.yas.nhs.uk/AboutUs/YAS_in_the_NHS.html`).

General Practitioners are the primary point of contact for patients and are considered as the 'Gatekeepers' to the NHS. With the exception of a few minority groups such as prisoners and the armed forces, each resident is allowed to register with a General Practice in the UK. Homeless people are 40% more likely not to be registered with General Practitioners when compared with the general population [52]. Patients may also be registered with both private and NHS General Practices. NHS General Practices provide a range of primary care services that include treatment of chronic and acute illness, prescribing of medication, referral to specialist/secondary care, preventative care such as screenings and immunisations, and health education such as smoking cessation, lifestyle advice and contraception.

General Practices keep lifetime medical records that include information such as patient demographics, 'signs, symptoms and diagnoses', primary care prescriptions (drugs and devices), immunisations, test results, referrals to specialist / secondary care, feedback from other care settings and lifestyle information (Body Mass Index (BMI), smoking, alcohol, exercise etc.). These General Practice records are primarily for General Practice use rather than for research. The quality, in terms of accuracy, completeness and detail, of these records varies between

practitioners, between practices and over time. General Practitioners are self-employed with contractual arrangements with the NHS and performance related pay. Introduced in 2004 the Quality and Outcomes Framework (QOF) is a payment management and payment system, which rewards practices for implementing good practice [116]. QOF forms part of the general medical services contract, which is voluntary but almost every General Practice takes part. QOF involves multiple best practice criteria, the level of adherence to which translates into level of payment to practices. Since the introduction of the QOF in 2004 data quality in General Practice medical records has improved.

2.5 Study Design & Data Sources

Prognostic studies are inherently longitudinal in nature. They are usually carried out in groups or cohorts of patients, who are followed over time allowing for an event of interest (outcome) to occur. The cohort is defined by certain criteria, know as selection criteria, such as the presence of one or more particular characteristics, e.g. having a certain disease, living in a certain geographic location, or being a certain age. There are several types of cohort studies can be used for prognostic modelling, including but not limited to retrospective cohort studies, prospective cohort studies, registry data, and nested case-control studies.

2.5.1 Study Design

In a prospective study, we design the study in advance to achieve some objective, and collect the data over time according to the study design or protocol. Once complete the data can then be used for analysis. The investigator is said to age with the study population (hence the term "prospective study" [305]). Using this design, we can better check specific selection criteria and use clear and consistent definitions of predictors and outcome, and recored them at predefined time points. Randomised, placebo-controlled clinical trials are a prominent example of prospective design, considered the 'gold standard' for assessing the safety and effectiveness of therapy [292]. They are designed to answer very specific questions about a particular treatment strategy, disease mechanism and patho-physiological process. In term of data quality, where

feasible, retrospective cohort studies are therefore generally preferable to retrospective studies. Prospective cohort studies are sometimes solely set-up for prediction modelling, but a more common design is that prediction research is done in data from Randomised Clinical Trials (RCT), or from prospective before-after trials [286]. The strengths are in the well-defined selection of patients, the prospective recording of predictors, usually with quality check, and the prospective assessment of outcome. However there are a few key limitations that may preclude its use for prognostic research.

A key limitation of prospective studies such as RCTs, may be in the selection of patients. Trials by their very nature select a very specific group of individuals, typically without comorbidity and/or polypharmacy. Typically stringent selection criteria are used, which may introduce bias, limiting the generalisability of any models developed on such data. Multi-centre trials help increase generalisability of findings, but are still contained to same limitation in the study design. Another challenge is feasibility, the high costs associated with trials often lead to trials that have relatively short duration and relatively small sample size. Insufficient study duration can be a problem in some prognostic studies, for example if we are looking to predict the 10-year risk CVD then data from a prospective that lasted 3 years would be of limited value. Sample size is key problem for prognostic research and has been discussed previously (section 2.8.1), Finally, there may also be ethical limitations on prospective designs, such a comparison group that didn't receive any treatment for a severe disease. It would unethical to intentionally deny patients vital treatment, however this may have been observed in patient records which could be achieved through a retrospective cohort study design.

The most common type of prognostic studies are of the retrospective cohort design, where patients are identified from patient records within a certain date ranges. These patients are then followed over time for the outcome, but the investigator looks back in time (hence the term "retrospective study" [305]). Strengths of a retrospective design include its simplicity, feasibility and relative cost effectiveness. These strengths are realised by the ease with which existing patient records can be searched. The low cost allows for relatively long time horizons and large sample sizes, which can be important for prognostic research, which are often prohibitively expensive in other study designs. The fact that the data is routinely collected is another strength, as this 'real-world' data does not suffer from the same bias that can exist in studies that are

prospective in nature. This is because the selection criteria in typically not as stringent and the data is observations from routine practice, generally leading to representative sample and thus a model and findings with higher generalisability to the target population.

Limitations include correct identification of patients, reliable recording of predictors and outcomes, and to a lesser extent, sample size. Challenges arise in the correct identification of patients, which has to be done in retrospect. If certain selection criteria are poorly defined, ill conceived or insufficient, or some of the information is missing or incorrectly recorded, this may lead to selection bias. Also the recording of predictors has to have been reliable for use in prediction modelling. Similarly, the outcome has to reliable. This may be relatively straightforward for an outcome such as all-cause mortality in cohort that typically would die in hospital, which may be reasonably well recorded in hospital patient records. However, if we are interested in death attributed to certain disease or cause, especially if often occurs outside of hospital, then the outcomes may need to be verified using additional information from other sources such the national statistics bureaus like the Office for National Statistics in the UK. Finally, sample size may be limited in single-centre studies, however this is not usually a problem in collaborative multi-centre studies. Retrospective cohort studies are often considered inferior to randomised trials, as, in some cases, they have been shown to overestimate treatment effects [292]. Conversely however, when RCT data has been used for prognostics modelling, they often have poor performance in groups of patients that differ from the very specific group used in the trial. Despite this, retrospective cohort studies, if well designed and conducted appropriately, can be a valuable and effective approach to determining associations between specific exposures and outcomes, and are the method of choice when it is not possible to conduct randomised trials [292]

2.5.2 Electronic Patient Records

The strengths of the retrospective design are further enhanced by the increasing availability of Electronic Patient Records (EPR) in primary and secondary care. The electronic recording clinical patient data has come a long way since its inception during the 1980s. In the UK alone, there are three predominant systems used both for research and for clinical records: the CPRD [39], The Health Improvement Network (THIN) [294] and QResearch [246, 274].

Since the 1990s the use of EPRs in the UK has increased, primarily as a result of UK government initiatives including the provision of financial incentives for general practice to develop electronic clinical registers (i.e. QOF) and through the proposed implementation of a single NHS computer system for EPRs (National Programme for Information Technology, NPfIT). The UK has a strong reputation of producing quality EPR-based research through publication of numerous studies, both in the validity and quality of the databases themselves, and in prognostic and epidemiological research performed using retrospective cohorts developed from their data [274]. These studies have confirmed the high quality of the data recording in these primary care databases [157, 289, 296]. Whilst clinical coding practice remains an area from improvement [100], this has been mitigated to some extend by the linkage with registry, secondary care and laboratory data [157, 288, 104].

There are several advantages of basing research on a national system of networked data. Firstly, the use of medical data collected across the UK provides a broadly representative sample. This is realised by the broad geographical spread individual practice databases across the UK allowing for a broad generalisation of the UK population. Secondly, many patients have complete data available form 1980, providing an excellent source of longitudinal data for prognostic research, in some cases providing heath information over the patents lifetime. Finally, the size (number of patients) of the databases not only enables more complicated question to be answered with fewer assumptions, but also enables the research into rarer diseases with low incidence rates, that couldn't be feasibly studied any other way. The CPRD observational dataset alone consists of longitudinal, anonymous records from 681 primary care practices and over 15 million patients throughout the UK (based on the March 2014 release) [39], with a similar number of practices (754) and patients ($>$13m) in QResearch's EMIS database [246].

Due its merits of feasibility, large sample size and high generalisability, a retrospective observational cohort design is the preferred study design for experiments conducted for this thesis. The data source proposed for development of such a cohort is routinely collected data from UK general practice patient records, linked with other data sources such as secondary care data and Office of National Statisitcs (ONS) data. This will be achieved using the CPRD (www.cprd.com). This decision to used CPRD as apposed the other primary care databases is that Cardiff School of Medicine already has licensed access to CPRD and significant experience

of using CPRD for high quality research.

2.6 The Clinical Practice Research Datalink

The CPRD (formerly the General Practice Research Database, GPRD) is a computerised database of anonymised longitudinal medical records collated from UK General Practitioners working in primary care [39]. General Practices using Vision Patient Administration Systems [133] can contribute data to the CPRD. Practices are paid for contributing to the scheme based on number of patients and quality of data provided. CPRD is considered by many as the GOLD standard [39], containing 15.8 million patient records collected continuously since 1987. At the time of writing, information on approximately 4.8 million active (alive) patients in the UK, equivalent to about 7% of the UK population, are collected from 681 general practices nationwide as detailed in figure 2.2 below.

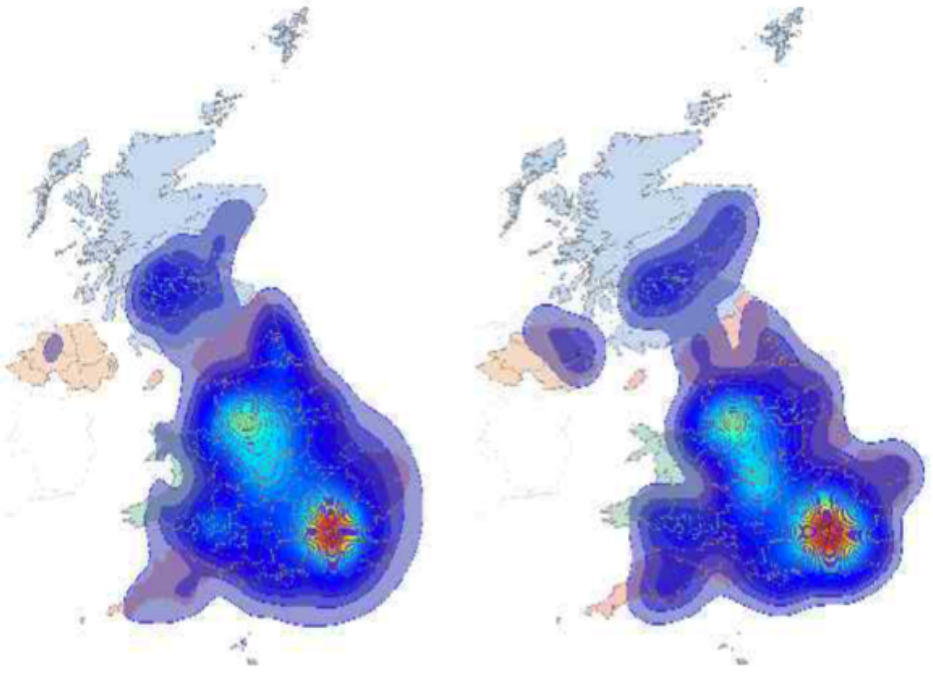


Figure 2.2: Heat map depicting smoothed estimates of population density (left) and CPRD coverage (right) in the UK.

CPRD also has an increasing number of links to secondary care such as Hospital Episode Statistics (HES), registries and death data. Information collected by CPRD along with some examples are detailed in Table 2.1 below. The UK Medicines and Healthcare products Regulatory Agency (MHRA) run the CPRD on a non-profit making basis. The MHRA is an Executive Agency of the DoH. It fulfils a critical public health role in the UK by ensuring that all medications and medical products meet appropriate standards of safety, quality and efficacy.

Table 2.1: Types of information held in CPRD

Information	Examples
Demographics	Smoking and drinking status, exercise, etc. Age, gender, height, weight
Medical symptoms/diagnosis and comments	Historical diagnosis Historical diagnosis Hospital referrals
All recorded prescriptions (drugs and devices)	Strength Dosage
Referrals	Hospitals or specialists
Registration details	Dates Status e.g. transferred out
Status e.g. transferred out	Immunisations Test results Consultations Repeat prescription schedules

2.6.1 Clinical Practice Research Datalink Governance

Established in 2006 by the Secretary of State, the Independent Scientific Advisory Committee (ISAC) is an independent non-statutory advisory body nominated to review the scientific merit of proposals for research using data from CPRD and safeguard patient confidentiality. ISAC approval is required for all studies using CPRD data, where destined for publication or for communication with a third party. ISAC approval is also required for any study that intends to

use data from a CPRD data linkage scheme.

2.6.2 Data Model

CPRD data is made available through online tools and static databases version created each month. The CPRD data model illustrated in figure 2.3 consists of the following data tables:

Patient demographics including age, sex

Practice region and collection information

Consultations duration and linkage

Therapy prescribing from general practice

Immunisations vaccination details

Staff General Practitioner, nurse, locum

Clinical medical history and diagnosis

Additional Clinical Details additional information from structured data areas

Referral information about secondary care

Tests structured numerical and qualitative lab results

The Patient table contains basic patient demographics and registration details for the patients. The Practice table contains details of each practice, including region and collection information. The Staff table contains practice staff details, with one record per member of staff. The Consultation table contains information relating to the type of consultation as entered by the General Practitioner from a pre-determined list. Consultations can be linked to the events that occur as part of the consultation via the consultation identifier (consid). The Clinical table contains medical history events. This file contains all the medical history data entered on the General Practitioner system, including symptoms, signs and diagnoses. This can be used to identify any clinical diagnoses, and deaths. Patients may have more than one row of data.

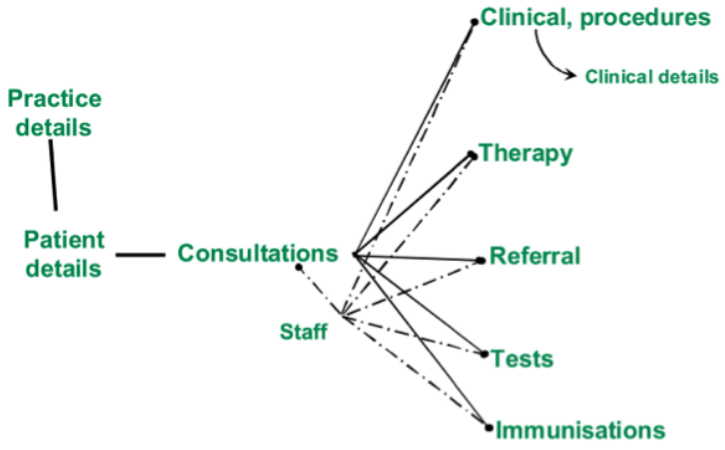


Figure 2.3: Figure depicting the CPRD data model

The data is coded using Read codes (section 2.7.1), which allow linkage of codes to the medical terms provided. The Additional Clinical Details table contains information entered in the structured data areas in the General Practitioner's software. Patients may have more than one row of data. Data in this file is linked to events in the clinical file through the additional details identifier (adid). The Referral table contains referral details recorded on the General Practice system. These files contain information involving patient referrals to external care centres (normally to secondary care locations such as hospitals for inpatient or outpatient care), and include speciality and referral type. The Immunisation table contains details of immunisation records on the General Practitioner system. The Test table contains records of test data on the General Practitioner system. The data is coded using a Read code, chosen by the General Practitioner, which will generally identify the type of test used. The test name is identified via the Entity Type, a numerical code, which is determined by the test result item chosen by the General Practitioner at source. There are three types of test records, involving 4, 7 or 8 data fields (data1 - data8). The data must be managed according to which sort of test record it is. Data can denote either qualitative text based results (for example 'Normal' or 'Abnormal') or quantitative results involving a numeric value. The Therapy table contains details of all prescriptions on the General Practitioner system. This file contains data relating to all prescriptions (for drugs and appliances) issued by the General Practitioner. Patients may have more than one row of data. Drug products and appliances are recorded by the General Practitioner using the Multilex product code system (section 2.7.2).

Full descriptions of fields in each table are provided in the Appendix [TBC]. All files can be linked using the encrypted patient identifier (patid). The last three digits of the patient identifier (patid), and staff identifier (staffid) denote the identifier of the practice (pracid) that the patient, or staff belongs to. The mapping column references information relating to the use of data in the field. It specifies lookup references, linkages to other tables, and information on decoding numerical values. A mapping of 'None' indicates the existence of raw data in the field. 1.

2.6.3 Data Quality

In large observational databases such as CPRD data quality varies between Practices and over time. CPRD uses data quality markers to ensure internal consistency of patient data, complete longitudinal records and complete, continuous, plausible practice level data. There are two principle data quality markers used in CPRD, a patient 'acceptability' flag and practice UTS ('up-to-standard') date. The patient 'accept' flag indicates whether the patient record is of an acceptable quality and integrity. Practice UTS dates refer to the date at which the practice is of research quality. CPRD data quality is driven by incentivisation, Medico-legal factors and feedback reports.

2.7 Clinical Coding

A clinical coding system is a coded thesaurus of clinical terms designed to enable users to make effective use of clinical computer systems. Accurate and consistent coding of clinical data onto the practice computer clinical system is a vital step in the move towards the Patient Electronic Health Record and consistent analysis [224]. Several clinical classification systems are relevant to this research project are described here.

2.7.1 Read Codes

The NHS mandates the use of the Read Clinical Classification Version 3 [221], more commonly known as Read Codes, in general practice. Read codes are a coded thesaurus of clinical terms

designed to enable General Practitioners, Practice Nurses and administration staff to make effective use of clinical computer systems. Read medical codes have been used in the VISION Patient Administration System since 1995 as a required field for a clinical entry into a patient record. The Read codes have a hierarchical structure made up of 5 levels of detail in the 5-character version (5 byte) of the codes and 4 levels in the earlier 4-character version (4 byte). The characters used are numbers 0 to 9, upper case letters A to Z and lower case letters a to z. The codes are case sensitive.

Table 2.2: Hierarchical structure of read code for Acute anteroapical infarction

Read Code	Term
G....	Circulatory system diseases
G3...	Ischaemic heart disease
G30..	Acute Myocardial Infarction (MI)
G301.	Other specified anterior MI
G3010	Acute anteroapical infarction

Table 2.3: Hierarchical structure of read code for Allotransplantation of heart and lung.

Read Code	Term
7....	Operations, procedures, sites
79...	Heart operations
790..	Heart wall, septum and chamber operations
7900.	Transplantation of heart and lung
79000	Allotransplantation of heart and lung

Tables 2.2 and 2.3 give examples of the read code hierarchy. The top level of the hierarchy is known as the chapter (chapters 'G' and '7' in examples above). Chapters starting with letters A-Z indicating referring to diagnosis codes and chapters starting with numbers 0 to 9 indicating process of care codes, in Table 2.2 'Diagnosis' and in Table 2.3 'Operations'. Each successive level of the hierarchy provides more detail to a concept.

2.7.2 International Classification of Diseases

The International Classification of Diseases (ICD) is a health care classification system that provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease [319]. Under this system, every health condition can be assigned to a unique category and given a code, up to six characters long. Such categories can include a set of similar diseases. The ICD is published by the United Nations-sponsored World Health Organisation (WHO) and revised periodically and is currently in its tenth revision ICD-10 which was implemented in 1992. The basic ICD is a single coded list of three-character categories, each of which can be further divided into up to 10 four-character subcategories. In place of the purely numeric coding system of previous revisions, the Tenth Revision uses an alphanumeric code with a letter in the first position and a number in the second, third and fourth positions. The fourth character follows a decimal point. Possible code numbers therefore range from A00.0 to Z99.9. Table 2.4 below describes the structure of ICD-10 code I21.0 for Acute transmural MI of anterior wall.

Table 2.4: Hierarchical structure of ICD-10 code for Acute transmural MI of anterior wall.

ICD-10 Code	Term
I....	Chapter IX: Diseases of the circulatory system
I2...	Ischaemic heart diseases (I20-I25)
I21..	Acute MI (I21.0-I21.9)
I21.0	Acute transmural MI of anterior wall

The classification is divided into 21 chapters (see table 2.5 below). The first character of the ICD code is a letter and each letter is associated with a particular chapter. Each chapter contains sufficient three-character categories to cover its content; not all available codes are used, allowing space for future revision and expansion. The chapters are described in table 2.5 below.

The desk reference for prescribing by General Practitioners in the UK is the British National Formulary (BNF) [28] provides information on the selection, prescribing, dispensing and administration of medicines. The BNF is arranged into chapters and sections, with some phar-

maceutical products being a member of more than one BNF chapter. General Practitioners contributing to CPRD code their selection, prescribing, dispensing and administration of pharmaceutical products using the Multilex drug knowledge base. Multilex is a proprietary drug terminology used by CPRD that holds clinical and commercial information on more than 75,000 pharmaceutical products and packs [82]. Products within Multilex include all branded prescription medicines, generic medicinal products, Pharmacy and General Sales List medicines, appliances included in the Drug Tariff, supplementary and specialist dietary foods, diagnostic and monitoring agents, homeopathic remedies and other NHS products available on prescription. UK practices and therefore CPRD use the Read Clinical Classification system, but often CPRD alone does not hold all the information required for research. In these cases we need to augment the data in CPRD with data from other sources, such as those provided by the linkage schemes (discussed later). Many of the additional datasets provided by linkage schemes use different coding systems, namely the ICD. The Read codes are cross-referenced to ICD-10, but not ICD-9. However there are challenges as the cross referencing is far from perfect, as in some cases there is not a one-to-one mapping between conventions. This is further complicated by the fact that some of data is coded using ICD-9, which does not always have a direct mapping to ICD-10 or Read. In these cases the expertise of clinicians is still required. Steps are being made to address this as the NHS National Programme for IT is using the SMOMED Clinical Terms system as the standard terminology in the NHS Care Records Service. SNOMED Clinical Terms (SNOMED CT) claims to be the most comprehensive, multilingual clinical healthcare terminology in the world and incorporates all Read (Version 3) terminology via 1:1 mapping [135].

The CPRD holds a wealth of information about primary care but often primary care research requires information that is not typically captured by general practice. In such cases CPRD data may need to be augmented with data from other sources. The MHRA aims to optimise CPRD research outputs by maximally linking person level data from different healthcare domains. Linkages that are available include:

- HES
- Hospital prescribing

- Cancer registries
- Myocardial Ischemia National Audit Project (MINAP)
- Office of National Statistics Mortality (death certificates)
- National Bone and Joint Registry (NJR)
- Socio Economic Status (Patient Level)

Data linkage is implemented on a practices level, with practices having the options whether or not to consent to the CPRD linkage scheme. CPRD has limited linked data available via its linkage scheme; linkage is only available to English practices consenting to participate in the scheme, representing 357 (71%) of the 500 English practices in CPRD and 4% of the 8324 practices in England [222]. The CPRD organisation is the custodian of the linked data and in most cases holds the full datasets. Figure 2.4 below illustrates a typical CPRD data linkage process. Data provided by these linkage schemes is not part of the base CPRD dataset and needs to be requested on study-by-study basis through the ISAC by submission and approval of an ISAC protocol. Granted ISAC protocols relating to the experiments in chapters 5 and 6 of this thesis are detailed in Appendices A and F, respectively.

As indicated above there are several data linkage schemes, those schemes that are directly relevant to this research, ONS Mortality, HES and Socio-Economic Status (SES) have been discussed in more detail below.

In the UK death registration is carried out by Local Registration Services in partnership with the General Register Office. All deaths must be reported to the Register General and sudden deaths require coroner involvement. The cause of death is recorded on Part I and Part II of the death certificate. Part I consist for three levels; 1(a)- Disease or condition directly leading to death; 1(b)- Other disease or condition, if any, leading to 1(a); 1(c)- Other disease or condition, if any, leading to 1(b). Part II details other significant conditions contributing to the death but not relating to the disease or death causing it. Cause of death is coded using the ICD coding system. ICD-9 has been used in the UK since 1979, with ICD-10 introduced in Scotland in 2000 and rest of the UK in 2001. Official death dates are available for 99.9% of English population and cause of death is only available since Jan 2001. ONS mortality data is not part of the

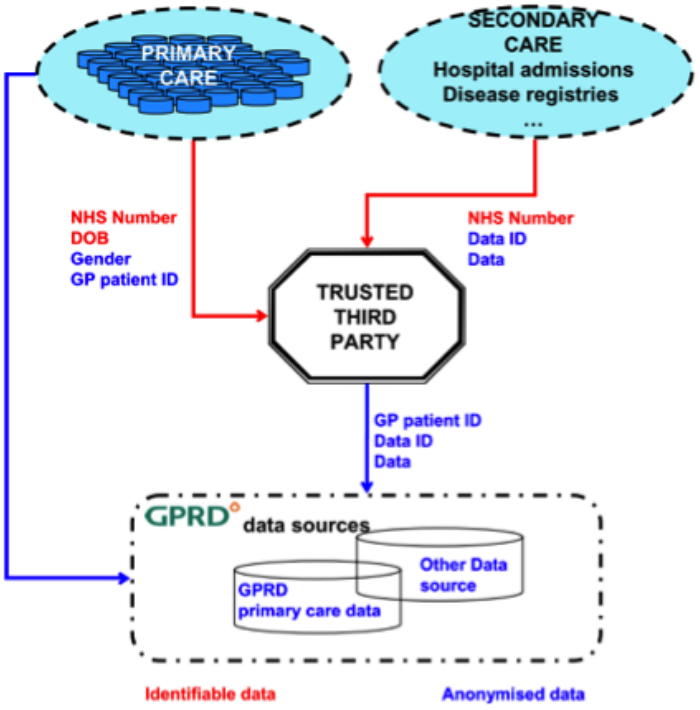


Figure 2.4: Figure depicting a typical CPRD data linkage process

base CPRD dataset, limited to English practices participating in the linkage scheme and needs to be requested on study-by-study basis through the ISAC. HES are a record-level database of hospital admissions and outpatient attendances for all NHS trusts in England [223]. HES holds admission data from 1989 and outpatient data from 2003. The HES dataset consists of patient, admission, discharge, clinical, speciality, critical care and maternity information. Clinical information is recoded for each episode, with diagnoses coded in ICD-10 and operative procedures coded using Office of Population Censuses and Surveys, Classification of Surgical Operations and Procedure (OPCS) Version 4 [220]. HES data is not part of the base CPRD dataset, limited to English practices participating in the linkage scheme and needs to be requested on study-by-study basis through the ISAC. CPRD has derived practice-based SES using the Index of Multiple Deprivation linked to the practice postcode for all four countries of the UK. SES is also available at ONS small area level (100 houses) using patient postcode. Two SES scores are available from CPRD; Townsend score and the Index of Multiple Deprivation. SES data is not part of the base CPRD dataset, limited to English practices participating in the linkage scheme and needs to be requested on study-by-study basis through the ISAC.

2.8 Linear Statistical Methods

Prediction is primarily an estimation problem [286]. For example, what is the risk of dying over tens years? But prediction is also about testing hypotheses. For examples, how does the management of blood sugars effect risk of death on patients with diabetes? Or more generally, what are the important predictors or risk factors associated with a particular disease. Statistical models may serve to address both estimation and hypothesis testing. In the medical literature much emphasis has traditionally been given to the identification of predictors.

Statistical models summarise patterns of the data available for analysis. In doing so, it is inevitable that assumption need to be made. Testing of underlying assumptions is especially important if specific claims are made on the effect of a predictor. Statistical models for prediction can be discerned into three main classes: regression, classification and description. Hereafter, these will be jointly referred to as linear statistical methods or models. Regression models are the most widely used statistical models in the medical domain [286]. Statistical modelling to make predictions encounters various challenges, including dealing model uncertainty and limited sample size.

2.8.1 Model Uncertainty & Sample Size

Model uncertainly arises from the fact that we usually do not fully pre-specify a model before we fit it to a dataset [36, 73]. Often, an iterative process is followed with model checking and model modification. On the other hand, standard statistical methods assume that a model was pre-specified. In that case, parameter estimates such as regression coefficients, their corresponding confidence intervals, and p-values are largely unbiased [286]. Whenever some part of the data is used to inform the structure of the model in some way there is potential for bias to occur, and for underestimation of the uncertainty of the conclusions drawn from the model. Fortunately, model uncertainty has been actively researched and as a result some statical tool are available to help study model uncertainty. Statistical resampling methods, namely *bootstrapping* help with uncertainty during the development and validation of statistical models.

Sample size is another key challenge in statistical modelling. A sufficient sample size is import-

ant to assures any scientific question with empirical data. We have to realise that the effective sample size may often be much smaller than indicated by the total number of subjects in a study [112]. For example, if we have relatively large number of patients but are interested in a relativity rare event with low incidence rate, then it is this small number events that is the effective sample size. Large sample size facilitates many aspects of prediction research. For multivariate prognostic modelling, a large sample size allows for selection of predictors with simple automatic procedures such as stepwise feature selection methods (although the value of such methods is topic of much debate) and reliable testing of model assumptions. Conversely, with a small sample size we have to be prepared to make stronger modelling assumptions. When the sample size is very small we can only ask relatively simple questions, while more complex questions can be addressed with larger sample sizes [286]. Therefore the ambitions of the research questions may need to be tempered with the effective sample size of the data available.

2.8.2 Survival Analysis

Suppose that we wanted to study the occurrence of a particular event (or events) in a cohort of patients. If the time until the event is not of interest, but rather just whether the event occurred or not, then response could be represented as a simple binary outcome and modelled using logistic regression. For example, in analysing post- surgery morality, it may not be important whether patients die 30 days after or two years after the procedure, just that that they perish. However, typically with long-term and chronic conditions in domains such as cardiology oncology and endocrinology, the time until event is important. For example, in a study of CVD in patients with diabetes, whether a patient experienced a cardiovascular event 6 months, or 6 years after diagnosis ofT2DM is very different. An analysis that simply counted the number of events would be discarding valuable information and sacrificing statistical power [112].

Generally, survival analysis is collection of statistical procedures for the analysis of data in which the outcome of interest is time until event, which is often called *failure time*, *survival time*, or *event time*. More generally, this type of outcome or response is referred to as a *survival outcome*, which different to and not to be confused with a *survival endpoint*. The latter is a general term used in indicate that the endpoint or event of interest is death, but not necessarily

that it is time until death that is of interest. Whereas, a survival outcome, survival analysis or use of a survival model implies that the outcome of interest is time to event only - the event itself may or may not be survival in the wider sense (i.e death). For example, one may use a binary survival endpoint (dead/alive) and logistic regression to model post-surgery survival. Conversely, one may use a survival analysis to model time to non-fatal MI, which would be using a survival model and thus employing a survival outcome, but in fact have nothing to do with survival (mortality) in the more general sense. Examples of survival outcomes of interest include time until cardiovascular death, time until death or MI, or time until incidence or recurrence of hypertension. More specifically, survival analysis involves the estimation of the distribution of the time it takes for an event to happen to a patient based some set of features, which are also known as *explanatory variables*, *predictors*, *risk factors*, *features*, or *covariates*.

Censoring

A key characteristic of survival data is that the follow-up of patients is typically incomplete [286]. Survival analysis must address a key analytical problem called *censoring*. For example some patients may have been followed 5 days, some for 15 days, etc., and were are interested in predicting 30-day survival. Patients with such incomplete data are said to be censored, often referred to as censored observations or censored data. In essence, censoring occurs when we have some information about an individuals survival time, but we don't know the survival time exactly in all subjects.

Using a simple example of censoring from Kleinbaum & Klein [162], consider leukaemia patients followed until they go out of remission, shown in figure 2.5 as **X**. If for a given patient, the study ends while the patient is still in remission (i.e. doesn't get the event), then the patient's survival time is considered censored. We know that for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission (i.e experiences the event) after the study ends, we do not know the complete survival time.

According to Kleinbaum & Klein, 2005 [162], there are generally three reasons why censoring occurs:

1. a person does no experience the event before **the study ends**;

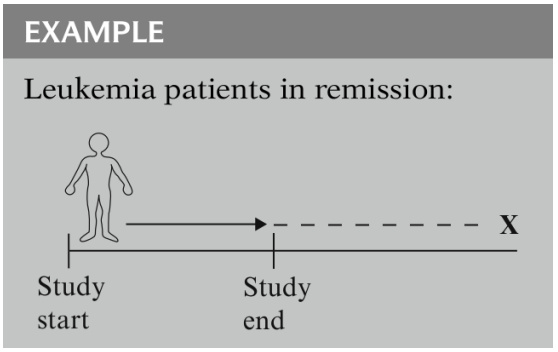


Figure 2.5: Example of censoring in leukemia patients followed until they go out of remission (source: Kleinbaum & Klein, 2005 [162]).

2. a person is **lost to follow-up** during the study period;

3. a person **withdraws from the study** because of death (if death is not the event of interest) or some other reason (i.e. adverse drug reaction or other competing risk)

These situation are graphically illustrated in figure 2.6. The graph describes the experience of several persons followed over time. An **X** denotes a person who got the event.

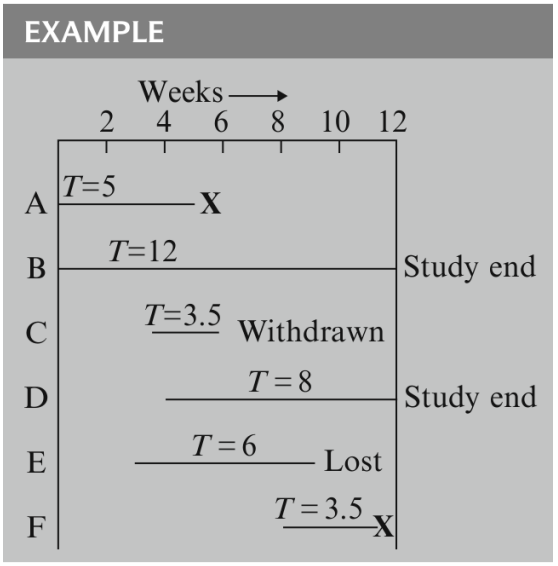


Figure 2.6: Example of different censoring in several leukemia patients followed over time (source: Kleinbaum & Klein, 2005 [162]).

Notation, Survial, and Hazard Functions

The two most popular quantitative terms considered in any survival analysis are the survivor (or survival) function, denoted $S(t)$, and the hazard function, denoted $\lambda(t)$. The survival function in the probability that an individual remains event-free longer than some specified time t. The hazard function $\lambda(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has remain event-free up to time t [162]. The hazard function is sometimes referred to as a conditional failure rate and in contrast to the survival function, which focuses on not experiencing the event, the hazard focuses on the event occurring. It is important to note that, although it may be helpful to think of the hazard as an instantaneous probability, it is not a probability as it can take on values greater than one [6]. Thus in some sense the hazard function can be considered as giving the opposite side of the information given by the survival function [162]. The distribution of time to a specific event dependent on a set a features or predictor variables $X = \{x_1, x_2, \ldots, x_k\}$, can be represented by five closely related functions given below:

$f(t|X)$ density function (pdf)

$F(t|X)$ cumulative distribution function (cdf)

$S(t|X)$ survival function

$\lambda(t|X)$ hazard function

$\Lambda(t|X)$ cumulative hazard function

As with other regression models, X can represent a mixture of binary, categorical, continuous, spline-expanded, and even ordinal predictors. These interrelated functions can expressed in Equations 2.1 - 2.5 [112, 162, 45, 6], where T denotes the response variable (usually time to event), t represents a specified time point, and X refers to a vector of features or explanatory

variables:

$$f(t,X) = \frac{\partial F(t,X)}{\partial t} = -\frac{\partial S(t,X)}{\partial t} = \lim_{\Delta \to 0} \frac{Prob\{t \leq T < t+\Delta, X\}}{\Delta} = \tag{2.1}$$
$$\lambda(t,X) exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \lambda(t,X)S(t,X)$$

$$F(t,X) = Prob\{T < t, X\} = \int_0^t f(u)\,\mathrm{d}u = 1 - S(t,X) \tag{2.2}$$

$$S(t,X) = Prob\{T > t, X\} = \int_t^\infty f(u)\,\mathrm{d}u = exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \tag{2.3}$$
$$1 - F(t,X) = exp[-\Lambda(t,X)]$$

$$\lambda(t,X) = -\frac{\partial log S(t,X)}{\partial t} = \frac{f(t,X)}{S(t,X)} = \tag{2.4}$$
$$\lim_{\Delta \to 0} \frac{Prob\{t \leq T < t+\Delta | T > t, X\}}{\Delta}$$

$$\Lambda(t,X) = \int_0^t h(u)\,\mathrm{d}u = -log S(t,X) \tag{2.5}$$

Of the functions considered thus far, the survival function is the most intuitive as it describes the survival experience in the cohort. Whilst being the less intuitive, the hazard function in important for several reasons. Firstly, the hazard function is a measure of instantaneous potential, whereas the survival function is a cumulative measure over time. Secondly, it can be useful when fitting a model to data, specifying the specific model form for parametric modelling. Thirdly, it is the vehicle by which the mathematical modelling is performed, that is, most survival models are expressed in terms of hazard. Whichever quantity is most appealing, the key point is that by specifying and one of the probability density function, survival function, or hazard function allows the other two function to be ascertained by using the formula 2.1-6.

Maximum Likelihood Estimation

Whereas unknown parameters in linear regression are estimated using least squares estimation, in logistic regression and survival analysis they are estimated using *Maximum Likelihood Estimation (MLE)*.

Maximum Likelihood estimates of model parameters are derived by maximising the a likelihood function. The likelihood function is a mathematical expression that describes the joint

probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters [162].

The natural logarithm of the likelihood, the *Log Likelihood (LL)*, is normally used for convenience in numerical estimation. The LL is calculated as the sum over all subjects of the distance between the natural log of the predicted probability p for the binary outcome to the actual observed outcome y:

$$LL = \sum y \times log(p) + (1 - y) \times log(1 - p) \tag{2.6}$$

where y refers to the binary outcome and p to the predicted probability for each subject [286]. A perfectly fitting model would have an LL of zero.

Kaplan-Meier Estimator

As the true form of the survival distribution is seldom known, it is useful to estimate the distribution without making any assumptions [112]. When censoring is present, $S(t)$ can be estimated using the Kaplan-Meier [147] product-limit estimator, a nonparametric method for survival data, based on conditional probabilities. The KM methods deals with censored data, and provides attractive graphs on the relationship between predictor values and the outcome over time [286]. Also, differences between survival curves can be tested statistical significance using the log-rank test.

The product-limit estimator is a nonparametric maximum likelihood estimator [112]. The formula for the KM product-limit estimator of $S(t)$ is as follows. Let k denote the number of failures in the sample and let $t_1, t_2, \ldots, t_k$ denote unique event times (ordered for ease of calculation). Let d_i denote the number of failures at t_i and n_i be the number of subjects at risk at time t_i; that is, n_i = number of failure/censoring times > t_i. The estimator is then

$$S_{KM}(t) = \prod_{i:t_i<t} (1 - d_i/n_i) \tag{2.7}$$

The KM method is often used in prognostic modelling, either to perform univariate analysis by generating survival curves for various subgroups (e.g. females versus males) to investigate the effect of predictors on outcome, or to calculate the estimated proportions of subjects that

remain event-free at a certain time points. However, the KM approach is limited as it cannot handle continuous predictors, not can it consider the effect of multiple explanatory variables simultaneously. Because of censoring, logistic regression (a binary variable) is inappropriate. One could attempt to apply linear regression on event time, but again censoring usually makes such an analysis meaningless [286]. To overcome this limitation, several regression modelling approaches, that allow for multiple predictors, have been proposed to enable prediction of time to event in the presence of censored data [287]. Most of these models have come from the long-established statistical literature such as parametric survival models, including the Weibull, lognormal and Gompertz models and the semi-parametric Proportional Hazards model proposed by Cox [48].

Parametric Survial Models

The nonparametric KM estimator of $S(t)$ is a very good descriptive statistic for displaying survival data. However, for the purposes of prognostic modelling, we need to make more assumptions to allow the data to be modelled in more detail. In this section we discuss a class of survival models, called parametric models, in which the distributions of the outcome (i.e. time until event) is specified in term of unknown parameters [162].

Examples of parametric models commonly used in biomedical research include linear regression, logistic regression, and Poisson regression. With these models, the outcome (i.e. time until event) is assumed to follow some known distribution, such as the normal, binomial, or Poisson distribution. What is typically meant is that the outcome is assumed to follow some family of distributions with unknown distributional parameters that need to estimated. For example, if we assume an outcome follows a normal distribution, then we know the distributional family but not the specific or exact distribution. To achieve that we need to estimate the distributional parameters, typically a *shape* and *scale* parameters, which in the case of the normal distribution would be mean (μ) and variance (σ). For parametric regression models, the unknown distributional parameters are often estimated from the data, and sometimes the appropriateness of the choice of distributions family (i.e. the fit to the data) is evaluated using the data.

A parametric survival model is a model in which the outcome (i.e. time until event) is assumed to follow some known distribution. Parametric survival modelling requires choosing one or more distributions. Common distributions include the Weibull, Exponential (a special case of Weibull), log-normal, log-logistic, Gompertz, and generalised gamma distributions. Due to their specific properties, some distributions can accommodate only one type survival regression formulation (such as the Proportional Hazards (PH) or Accelerated Failure Time (AFT) specification, discussed in the next sections), whereas other distributions may accommodate multiple formulations. Machine Learning (ML) is used to estimate the unknown parameters of $S(t)$ (see section 2.8.2).

Providing the parametric form is correctly specified, there are several advantages in specifying a functional form for $S(t)$ and estimating any unknown parameters in this function (i.e. parametric survival modelling) . Firstly, concise and parsimonious equation and smooth estimates of $S(t)$, $\Lambda(t)$, and $\lambda(t)$, which can be especially useful in prediction modelling. Secondly, unlike non- and semi-parametric models, fully parametric models enable robust estimation of expected event times, typically by extrapolation. Due to their nonparametric nature, predictions made by the KM estimator and Cox regression (discussed in section 2.8.2) towards the end of follow-up are quite unstable, whereas with parametric models they are more robust, as survival is completely specified by a smooth function over the entire time horizon. In fact, from a technical view point these approached are only valid for time points in the data where an event occurred. Thirdly, due to the fully specified functional form, selected quantiles of the survival distribution can be easily computed. Finally, assuming the parametric form is correctly specified, more precise estimation of $S(t)$, when compared with nonparametric estimates such as $S_{KM}(t)$.

Parametric Proportional Hazards Models

The most widely used survival regression specification assumes that the effects of the covariates $exp(\beta X)$ are multiplicative (i.e. proportional) with respect to hazard $\lambda(t)$, know as the *proportional hazards assumption (PHA)*:

$$\lambda(t, X) = \lambda(t)exp(\beta X) \tag{2.8}$$

This regression formulation is called the *proportional hazards (PH)* model. Analogous to the PH (parametric and semi-parametric) model for a binary outcome in uncensored data, where we know whether or not the patient experienced the event in the time horizon of interest, is the logistic model. Multivariable logistic regression model is is the most widely used statistical technique nowadays for binary medical outcomes [107, 286, 307]. The PH model is the natural extension of the logistic model to the survival setting [286]. Indeed, the the PH model is equivalent to conditional logistic regression, with conditioning at times where events occur [184]. In the logistic model, we use an intercept in the regression effect, while in the PH model uses what is referred to as the *baseline hazard function*, *underlying hazard function*, *average risk profile*, or *hazard function for a standatd object*.

The $\lambda(t)$ part of $\lambda(t, X)$ is the baseline hazard function, which is a subject (i.e patient) with $\beta X = 0$. Any parametric hazard function can be used for $\lambda(t)$, and as we will see in the next section, $\lambda(t)$ can be left completely unspecified without sacrificing the ability to estimate β, by the use of the semi-parametric *Cox Proportional Hazards* model (section 2.8.2) [48]. As with other regression models, the vector of predictor variables $X = \{x_1, x_2, \ldots, x_k\}$ can represent a mixture of binary, categorical, continuous, spline-expanded, and even ordinal predictors. The effects of the covariates, the regression effect $\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ is modelled in much the same way as other settings. Depending on whether the distribution of the underlying hazard function $\lambda(t)$ has a constant scale parameter, βX may or may not have an intercept β_0. In multiple linear regression, the regression effect βX can be thought of as an increment in the value of the expected response Y. In binary logistic regression, βX specifies the log odds that $Y = 1$, or $exp(\beta X)$ multiplies the odds that $Y = 1$. Depending on whether the baseline hazard function $\lambda(t)$ has a constant scale parameter, βXmay or may not include an intercept β_o.

The PH model can also be rewritten in terms of the cumulative hazard and survival functions:

$$\Lambda(t, X) = \Lambda(t) exp(\beta X) \tag{2.9}$$

$$S(t, X) = exp[-\Lambda(t) exp(\beta X)] = exp[-\Lambda(t)]^{exp(\beta X)} \tag{2.10}$$

$\Lambda(t)$ is an "underlying" cumulative hazard function. $S(t, X)$, the probability of remaining

event-free up to time t, given the values of predictors X, can also be written as

$$S(t, X) = S(t)^{exp(\beta X)} \tag{2.11}$$

The regression effect βX is usually centred at the mean values of the predictors and the term $exp(\beta_k)$ is called a *relative hazard function* or a *hazard ratio*, which is the ratio of hazard of predictor k compared with the baseline hazard (i.e. average risk profile). The *hazard ratio* is similar to the odds ratio in logistic regression and arguably the most well known (and often misunderstood) term in survival analysis, in many cases is the function of primary interest as it describes the (relative) effects of the predictors.

Note that the regression effect relates the log hazard or log cumulative hazard. In the general regression notation, the PH model can be linearised with respect to βX, allowing distributional and regression parts to isolated and checked, using the following identities.

$$\begin{aligned} log[\lambda(t, X)] = \; & log[\lambda(t)] + exp(\beta X) \\ & log[\lambda(t)] + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \end{aligned} \tag{2.12}$$

$$\begin{aligned} log[\Lambda(t, X)] = \; & log[\Lambda(t, X)] = log[\Lambda(t)] + exp(\beta X) \\ & log[\Lambda(t)] + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \end{aligned} \tag{2.13}$$

Accelerated Failure Time Models

Besides the PH survival regression formulation, where effect of predictors is multiplicative with respect to hazard, other regression formulations can be specified. The *AFT model* is commonly used; it assumes that the effect of covariates are multiplicative (i.e. proportional) with respect to event time or additive with log event time, known as the *AFT assumption*. The effect of the predictor is alter the rate at which the subjects proceed along the time axis (i.e. to accelerate or shorten the time to event [143]). The model is

$$S(t, X) = \psi \left(\frac{log(t) - \beta X}{\sigma} \right) \tag{2.14}$$

where ψ is any standardised survival distribution function. The parameter σ is called the *scale parameter*. The model could also be stated as $\psi \sim [log(T) - \beta X]/\sigma$. The Weibull and

exponential are the only two distributional families that can accommodate both AFT and PH assumptions.

The interpretation of parameters in different in AFT models from that of PH models. The AFT assumption is applicable for the comparison of times until event. Whereas, the PH assumption is applicable for the comparison of hazard. Many parametric survival models are AFT models, rather than PH model.

Using an example from Kleinbaum & Klein [162], the AFT assumption can be illustrated by considering the comparison of survival functions among smokers $S_1(t)$ and non-smokers $S_2(t)$. The AFT assumption can be expressed as $S_2(t) = S_1(\gamma t)$ for $t \geq 0$, where γ is a constant called the *acceleration factor* comparing smokers and non-smokers. In a regression framework the acceleration factor γ could be reparameterised as $exp(\beta)$ where β is a parameter to be estimated from the data. With this parameterisation, the AFT assumption can be expressed as $S_2(t) = S_1[exp(\beta)t]$ or equivalently $S_2[exp(-\beta)t] = S_1(t)$ for $t \geq 0$.

The AFT assumption can also be expressed in terms of random variables for survival time rather than the survival function. If T_2 is a random variable (from some known distribution) representing the event time for non-smokers and T_1 is a random variable representing event time for smoker, then the AFT assumption can be expressed $T_1 = \gamma T_2$. The $log\lambda$ and $log\Lambda$ transformations of the PH model has the following equivalent for AFT models.

$$\psi^{-1}[S(t, X)] = \frac{log(t) - \beta X}{\sigma} \tag{2.15}$$

letting ϵ denote a random variable from the distribution ψ, the model is also

$$log(T) = \beta X + \sigma\epsilon \tag{2.16}$$

So the property of the response T of interest for regression modelling is $log(T)$. A one-unit change in X_j is then most simply understood as a β_j change in log event time. The one-unit change in X_j increases the event time by a factor of $exp(\beta_j)$ [112].

The acceleration factor is a key term of interest obtained from AFT models. It enables the evaluation of the effect of predictors on time until event, just as the hazard ratio allows the evaluation of the effect of predictors on hazards . In AFT models the acceleration factor describes the "stretching out" or contraction of survival functions when comparing one group to

another [162]. Hence the idea accelerating (or decelerating) toward the event of interest. More specifically, the acceleration factor is a ratio of survival times corresponding to any fixed value of $S(t)$ [162].

The idea is graphically illustrated by examining for Group 1 (G1) and Group 2 (G2) show in figure 2.7. For any fixed value of $S(t)$, the distance of the horizontal line from the $S(t)$ axis to the survival curve for $G = 2$ is double the distance to the survival curve for $G = 1$. Notice that the median survival time (as well as 25th and 75th percentiles) is double for $G = 2$. For AFT models, the ratio of survival times is assumed constant for all fixed values of $S(t)$ [162].

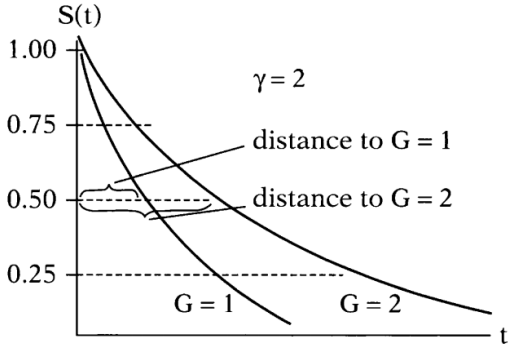


Figure 2.7: Graphical illustration of the AFT assumption and the acceleration factor γ in the comparison of the survival curves amoung two groups (source: Kleinbaum & Klein, 2005 [162]).

The Cox Proportional Hazards Model

In medical and epidemiological studies the Cox Proportional Hazards model [48] (or Cox regression) is the most is the most often used model for survival outcomes [112, 286]. Just as in the fully parametric form of the PH model, the Cox model is analogous to the logistic model for binary uncensored data, for a survival outcome. Indeed, the Cox model is equivalent to conditional logistic regression, with conditioning at times where events occur [184]. Again, just as is the case of the logistic model, different variants (some simpler and some more extensive) exist, which can be seen as special cases or extensions of the Cox model.

The Cox regression model is often stated in terms of the hazard function cite [307, 286]:

$$\lambda(t, X) = \lambda(t)exp(\beta X) \tag{2.17}$$

Note that the Cox model is an example of a PH model where we do not include an intercept parameter in the regression effect $\beta X = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k dfd$.

It as a semi-parametric method; it makes a parametric assumption concerning the effect of the predictors in the hazard function (i.e proportionality of effect during follow up), but baseline hazard function $\lambda(t)$ is non-parametric meaning that it makes no assumption regarding the nature of the baseline hazard function itself. The Cox PH model assumes that predictors act multiplicatively on the hazard function but does not assume that the hazard function is of a particular form, such as exponential or Weibull. This is an advantage of the model, as in many situations either the form of the true hazard function is unknown or it is complex. This is particularly advantageous when the primary the interest of the study is the effect of the predictors, rather than the shape of $\lambda(t)$ (which is often the case), as the Cox PH models allows the analyst to essentially ignore $\lambda(t)$. Another advantage is that the Cox PH model is less effected by outliers when compared with parametric methods, due way it used rank ordering of failure and censoring times.

For estimating and testing regression coefficients, the Cox model is as efficient as parametric models even when all the assumptions of the parametric model are satisfied [74]. When a parametric model's assumptions are not true (i.e. when a Weibull model is used and the distribution of event time does not in fact follow a Weibull distribution and thus the choice of model is incorrect), the Cox analysis os more efficient that the parametric analysis. For prognostic modelling we need to predict the risk of the event over time, for example by using the cumulative hazard or survival function. In exactly the same way as in with the general PH model formulation, the Cox PH can be rewritten in terms of the survival functions:

$$S(t, X) = S(t)^{exp(\beta X)} \tag{2.18}$$

However, in the Cox the non-parametric baseline survival hazard $S(t)$ has not been specified, it is usually estimated from data using the mean values of the predictors. The baseline survival is estimated from the non-parametric baseline hazard function as

$$S(t) = exp[-\Lambda(t)] \tag{2.19}$$

where $\Lambda(t)$ is the cumulative hazard at time t. The baseline survival in the training data determine the precise time points where we can make prediction for, which is not very natural for application of the model in new subjects.

Whilst Cox regression and the KM estimator are more flexible than parametric regression models in their dealing with the baseline hazard function, extrapolation is not readily possible with Cox or KM analysis because of their non-parametric nature. Predictions at the end of follow-up are quite unstable with Cox or KM analysis, and are more robust with parametric survival analysis. For estimation of the effect of predictors, the Cox model is often more suitable than an exponential or Weibull model. However, log-logistic models have been useful in situations where predictors worked especially during an early, acute phase of the hazard, which would show as non-proportional hazards in a Cox model [112]. Note finally that some of the more flexible methods for binary data have also been extended to survival models, but are not commonly used yet (e.g. neural networks) [115, 286].

2.9 Non-linear Statistical Methods

2.9.1 Evolutionary Computation

Other methods for survival analysis are based on Artificial Intelligence (AI) and machine learning. Hereafter, these will be jointly referred to as non-linear statistical methods or models. Conventional linear and non-linear regression techniques seek estimate (i.e. optimise) the parameters to some pre-specified model structure. In contrast, if a search process works simultaneously on both the model structure and model parameters, the technique is called *symbolic regression* [167]. Symbolic regression avoids imposing a priori assumptions, and instead infers the model from the data. To obtain the best quality approximation in symbolic regression we do not specify the size or structural complexity of the model in advance, and no particular model is provided as a starting point to the algorithm. Instead, initial solutions are formed by randomly combining mathematical building blocks such as mathematical operators. By not requiring a specific model to be specified, symbolic regression isn't affected by human bias, or unknown gaps in domain knowledge. It attempts to uncover the intrinsic relationships of

the dataset, by letting the patterns in the data itself reveal the appropriate models, rather than imposing a model structure that is deemed mathematically tractable from a human perspective. This approach has, of course, the disadvantages of having a much larger problem space. This space, referred to as the *search space*, comprises of all possible solutions to the problem at hand. In fact, not only is the search space for symbolic regression infinite, but also there are an infinite number of models that could perfectly fit a finite data set (providing its complexity isn't artificially limited in some way). Clearly the search space is clearly too vast for a blind random search. Therefore some intelligent adaptive way to search the space is required.

In computer science, EC is a subfield of AI and machine learning that borrows liberally from population biology, genetics and evolution. Algorithms chosen from this collection are known as Evolutionary Algorithms (EA), which are search techniques based on computer implementations of mechanisms inspired by biological evolution such as reproduction, mutation, recombination, and natural selection. The process of evolution by means of natural selection (descent with modification) was proposed by Charles Darwin in *"On the Origin of Species: By Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life"* [57] in 1859, to account for the variety of life and its suitability (adaptive fit) for its environment [29]. These evolutionary mechanisms describe how evolution actually takes place through modification and propagation of genetic material. Evolutionary algorithms are adaptive computational systems that utilise simplified versions of the processes and mechanisms of evolution, to search for (approximate) solutions to problems.

An EA is referred to as an Adaptive Strategy and a Global Optimisation technique, because it can effectively explore very large search spaces without being trapped in local optima. This makes EA well suited to symbolic regression, where the search space is vast. Also, because evolutionary algorithms require diversity in order to effectively explore the search space, the end result is likely to be a selection of high-quality solution, both in model structure and the corresponding set of parameters.

Most EA s may be divided into *generational* algorithms, which update the entire sample once per iteration, and *steady-state* algorithms, which update the sample a few candidate solutions at a time [195]. For many specific EA s, there are both generational and steady-state versions of the algorithm. Based on Luke 20013 [195], we present a basic generational form of an

evolutionary algorithm in algorithm 2.1. The basic generational EA first generates an initial starting set of solutions to the problem, know as an initial *population*, and then iterates through three procedures. Firstly, each of these solutions, known as *individuals*, are evaluated to assign a numerical measure of *fitness*, which is some measure of quality of the individual solutions ability to address the problem. Secondly, this fitness information is used to determine how to breed a new population of *children* form the current population who act as *parents*. Thirdly, the parent and children populations are joined somehow to form the next-generation population, and the cycles continues. These steps are repeated until a pre-specified stop criterion is satisfied, usually when a maximum number of generations is reached or when the best individual reached some defined level of quality (fitness).

Algorithm 2.1 An Abstract Generational Evolutionary Algorithm. [195]

1: $P \leftarrow$ Build initial population
2: $Best \leftarrow \square$ ▷ $\square$ means "nobody yet"
3: **repeat**
4: AssessFitness(P)
5: **for** each individual $P_i \in P$ **do**
6: **if** $Best = \square$ or Fitness(P_i) > Fitness($Best$) **then**
7: $Best \leftarrow P_i$
8: **end if**
9: **end for**
10: $P \leftarrow$ join(P, Breed(P))
11: **until** $Best$ is the ideal solution or we have run out of time
12: **return** $Best$

2.9.2 Genetic Programming

In its simplest form, *representation* is the data structure used to define the the solution, i.e. the individual. However, we can also think of representation as the approach we take constructing, modifying, and presenting the individual for fitness assessment. In EC there is no single representation of an individual. The representation is usually based on the type of problem we are

tying to address or personal preferences of the investigator. Chronologically we can distinguish the following main subclasses of EC as:

- Evolutionary Programming (EP) , was developed by Lawrence Fogel [87, 88] in the early 1960s. EP was originally based on graph representations (specifically finite-state automata). This family of algorithms are inspired by macro-level or the species-level process of evolution (phenotype, hereditary, variation) and is not concerned with the genetic mechanisms of evolution (genome, chromosomes, genes, alleles) [29].

- Evolutionary Strategies (ES) , introduced by Ingo Rechenberg and Hans-Paul Schwefel [249, 250, 163] in the mid 1960s, uses real-valued vectors, typically of fixed length, mainly for parameter optimisation. It is an almost identical approach to that of Fogel's EP which is also inspired by the same species-level process of evolution [29].

- Genetic Algorithms (GA) , invented by John Holland [128] in the 1970s, typically uses fixed-length bit-strings to encode solutions. The GA is inspired by population genetics (including heredity and gene frequencies), and evolution at the population level, as well as the Mendelian understanding of the structure (such as chromosomes, genes, alleles) and mechanisms (such as recombination and mutation) [29].

A fourth class of evolutionary algorithm was first proposed by Nichael Cramer in 1985 [50]. However, it is John Koza who is credited with the development and popularisation of the field of GP with though his considerable work and his 1992 monograph on GP [167], the seminal reference for the field. In his book Koza performs a number of GP experiments, evolving computer programs to solve a range of problems, including symbolic regression.

The most common form of GP, *tree-based GP* , uses trees as its representation. Consider the tree in figure 2.8, containing the mathematical expression $max(x + x, x + 3y)$. This is the *parse tree* of a simple program which performs this mathematical operation. There are several advantages to using parse trees to represent the solution, including, preventing syntax errors, which could lead to invalid individuals, and the hierarchy in a parse tree resolves any issues regarding function precedence. Tree-based GP features heavily in n Koza's seminal work in the field. However, in addition to trees, there are other important representations that include graph and linear representations [239, 17].

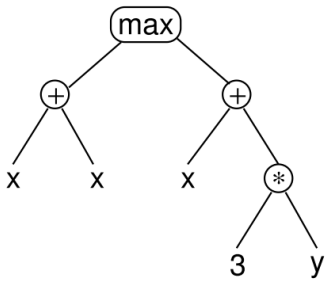


Figure 2.8: A symbolic regression parse tree representing $max(x + x, x + 3y)$ (source: Poli et al., 2008 [239]).

As with GAs, the GP algorithm is inspired by population genetics, and evolution at the population level, as well as the Mendelian understanding of the structure and mechanisms [29]. Genetic algorithms and GP are very similar EA s, their main difference is in their encoding of the solution, with the former is used in parameter optimisation, while the latter evolves the structure of the approximation model.

In GP we evolve populations of computer programs or solutions. That is generation by generation, GP stochastically transforms populations of programs into new, hopefully better, populations of programs (figure 2.9). GP automatically solves problems without requiring the user to know or specify the form or structure of the solution in advance [239]. This makes GP well suited to symbolic regression, where in addition searching for the solution to the complex associations between predictors and outcome, GP also searches for the optimal model structure. Which in turn makes GP well suited to prediction, primarily an estimation problem, where the mutual correlations between predictors and the outcome are to be estimated. GP has been shown to work well for recognition of structures in large datasets and has the intrinsic advantage of automatically select a subset of inputs or features during the evolutionary process [212].

Because GP has no fixed model structure it can represent complex non-linear associations that could not be achieved using linear regression techniques, and theoretically achieve higher predictive accuracy. However the flexibility of regression models can be greatly enhanced through the use of factional polynomial, restricted cubic splines and interaction terms, potentially increasing its predictive accuracy [22, 111, 112, 164]. However, it has be emphasised that this is not normally done as it complicate interpretation and the correct use of the appropriate regres-

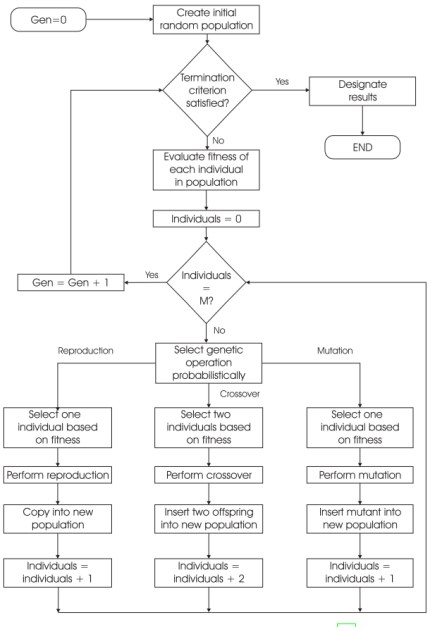


Figure 2.9: Generational GP flowchart (based on Koza [167]). M is the population size and Gen is the generation counter. The termination criterion can be the completion of a fixed number of generations or the discovery of a good-enough individual (source: Sipper., 2011 [278]).

sion methods is quite difficult, as they require extensive statistical knowledge [271]. Another characteristic of GP that is advantageous for symbolic regression is the automated selection of inputs. Not all input (i.e. predictor) variables have a significant effect in the outcome. Typically in clinical prediction modelling and survival analysis, we start out with a large set for candidate predictors and only a small subset if these are used in the final (i.e best) model. The task of identifying this subset is called *feature selection*, something that GP does inherently and robustly as part of the evolutionary process. The main drawback of GP for symbolic regression is its high computational cost, due to the potentially infinite search space. This can be attenuated to some degree by limiting the mathematical building blocks provided to the algorithm, based on existing knowledge of the problem domain and the system that produced the data. On the other hand, the recent availability of fast multi-core systems has enabled the practical application of GP in many real-world application areas.

The focus of this thesis is on regression models, which are the most widely used in the med-

ical domain, for the prediction cardiovascular risk in patients with T2DM. We only consider situations where there are a limited number of variables, say less than 25. This is in contrast to other research areas such as genomics (genetic effects), proteomics (protein effects), and metabolomics (metabolite effects). In these areas there often large number of candidate predictors (features), often $>$10,000. The application of GP to these types of biomedical problems is an active research area and experienced a degree of success (see chapter 3). Obtaining predictions form a model has to be separated from obtaining insights in the disease mechanisms and patho-physiological processes. Such insights are related to the estimated effects of predictors in a model. Often predictions models serve the latter purpose too, but the primary aim considered in this thesis is outcome prediction. However we endeavour to present and interpret and relationships discovered by the final GP system, effectively generating hypotheses on disease mechanisms and patho-physiological processes.

Our main focus is on the evolution of symbolic expressions (the clinical prediction models) and we will use a classical GP approach using trees to represent the models or solutions. A tree-based representation is adopted because it offers many benefits when implementing the GP algorithm in computer code. In many programming languages, particularly interpreted functional ones, expressions are internally represented as trees, meaning expression can be directly evaluated by the interpreter. This potentially enables the whole spectrum of functions available within a language to be used as inputs of building blocks to the GP system.

2.10 Motivation

The current body of research and a lack of consensus on which risk score is most appropriate for UK general practice suggests the need for cardiovascular risk models that are closely calibrated to the contemporary UK T2DM population. Such models would enable physicians and health policy makers to not only understand the risk of a certain outcome but also understand the changes in risk resulting from changes in treatment. We investigate the utility of tree-based GP for survival analysis, more specifically for the prediction of cardiovascular risk in UK patients with T2DM. To some, GP may not seem to be the best suited or obvious choice for clinical prediction modelling. Traditional statistic methods are well researched, widely taught,

widely available, relatively straightforward, and relatively fast to execute. However in addition to being widely used, they are widely abused, and the correct application of these traditional statistical models can often require significant statistical knowledge and expertise, as well specific knowledge of the application domain.

The main advantage of GP is that it performs a global search for a model, contrary to the local search of most traditional machine learning algorithms [76, 90] and statistical methods. This gives the potential to discover potentially complex non-linear patterns that could not be discovered by traditional techniques. Whilst GP is computationally expensive, the recent availability of fast multi-core systems had attenuated this some degree. Another advantage of GP is the fact that clinical prediction models can automatically generated without the, sometimes signifiant, statistical and clinical expertise required in traditional modelling approaches. Whilst GP is more complicated to validate by clinicians when compared to traditional statistical approaches, it is not a *black-box method* and can be validated, unlike some traditional machine learning algorithms. There have been several studies that have compared other novel machine learning algorithms with their classic statistical counterparts for survival outcomes [148, 138, 229, 248, 269, 265], however the results are mixed as to whether these machine learning methods offer improved performance. GP has been used in medical research for classification and, to a lesser extent, prediction. However its value for prediction on censored data, for survival analysis, is not yet been documented.

2.11 Summary Conclusions

We introduced clinical prediction modelling and its role and importance in predicting the risk of CVD in UK patients with T2DM. As discussed in section 2.4 the structure of the UK NHS means that UK General Practices are the 'gatekeepers' to the NHS, and as such primary care data, especially that which is linked with other sources of data, provides a rich resource for prognostic research. In section 2.5, we gave an overview of the aspects of design of prognostic studies, comparing and contrasting trials and observational studies. We outlined that trials by their very nature select a very specific group of individuals, typically without comorbidity or polypharmacy. The outcome of interest in this research project is CVD, which typically affects

older people who in turn typically have comorbidites and polypharmacy. We then argued that prediction of cardiovascular risk in patients with T2DM would be most appropriately evaluated using an observational cohort study. In sections 2.5.2 we detail how the strengths of retrospective study designs are further enhanced by the increasing availability of electronic patient records in primary and secondary care. In section 2.6 we described the CPRD and justified its selection as the suitable dataset for the experiments in this study, and in section 2.7 discussed the clinical coding conventions that we will to work with in the refining observational cohorts from CPRD.

In section 2.8, we discussed how time until event is an important outcome in long-term chronic conditions such as CVD and diabetes. This is because simply defining the outcome as whether or not the event happened, typically modelled using logistic regression, would be discarding valuable information and sacrificing statistical power. We also discussed a key characteristic of time until event data, censoring, and what specific challenges this presents when modelling such data. We then gave an overview of current linear statistical methods for addressing the challenges of censored data, a collection of methods referred to survival analysis. Cox Regression appears to provides the default framework for prediction of long-term chronic outcomes in the presence of censoring. Kaplan-Meier analysis provides a non-parametric method, but requires categorisation of all predictors. Parametric survival models are more complicated but are parsimonious and robust, and are are particularly useful for prediction at the end of, or even beyond the observed follow-up.

In section 2.9.1 we introduce AI and machine learning for survival analysis and give an overview of symbolic regression and the field EC, discussing its strengths and weaknesses. We discuss how symbolic regression has the potential to discover complex non-linear relationships between outcome and predictors, that could not be represented using classic linear statistical approaches.

In section 2.9.2 we introduce GP, a particular algorithm within the collection of techniques know as EC. In GP we evolve populations of models. That is generation by generation, GP stochastically transforms populations of models into new, hopefully better, populations of models We discuss how GP is a relatively recent technique that shows potential, which may improve the selection and transformation of predictors, and may lead to models with good predictive ac-

curacy in new patients. We discuss the relative strengths and weaknesses of GP. GP is well suited to symbolic regression, however its almost infinite search space make this a computationally extensive approach. However, this has been attenuated by the recent availability of fast multi-core systems.

Finally in section 2.10, we provided motivation for the thesis. The current body of research suggests the need for cardiovascular risk models that are closely calibrated to the contemporary UK T2DM population. GP is a recent approach that has shown potential when used in medical research for classification and, to a lesser extent, prediction. However its value for prediction on censored data, for survival analysis, is not yet been documented.

In the next chapter, we will focus on survey and critical assessment of existing research in relation to this thesis.

Table 2.5: Chapter structure of the ICD-10 classification system

Chapter	Blocks	Title
I	A00–B99	Certain infectious and parasitic diseases
II	C00–D48	Neoplasms
III	D50 –D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00–E90	Endocrine, nutritional and metabolic diseases
V	F00–F99	Mental and behavioural disorders
VI	G00–G99	Diseases of the nervous system
VII	H00–H59	Diseases of the eye and adnexa
VIII	H60–H95	Diseases of the ear and mastoid process
IX	I00–I99	Diseases of the circulatory system
X	J00–J99	Diseases of the respiratory system
XI	K00–K93	Diseases of the digestive system
XII	L00–L99	Diseases of the skin and subcutaneous tissue
XIII	M00–M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00–N99	Diseases of the genitourinary system
XV	O00–O99	Pregnancy, childbirth and the puerperium
XVI	P00–P96	Certain conditions originating in the perinatal period
XVII	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00–T98	Injury, poisoning and certain other consequences of external causes
XX	V01–Y98	External causes of morbidity and mortality
XXI	Z00–Z99	Factors influencing health status and contact with health services
XXII	U00–U99	Codes for special purposes

Chapter 3

Related Work

Next we focus on survey and critical assessment of existing research in relation to this thesis. First we review and assess the performance of the numerous existing cardiovascular risk scores for the general population and for patients with T2DM . Before reviewing previous work on GP in the biomedical and health domain, we review the wider field of machine learning for the survival analysis. Finally we assess previous work specifically on GP for prognostic research.

3.1 Cardiovascular Risk Scores for the General Population

In public health, prediction models may help target preventative interventions to subjects with relatively high risk of having or developing a disease. Various model have been developed to predict the future occurrence of disease in asymptomatic subjects in the population. Well known examples are the Framingham risk functions for cardiovascular disease [314]. The Framingham risk functions underpin several current policies for preventative interventions. For example, lipid-lowering therapies are only considered for those with relatively high risk of cardiovascular disease.

Since the Framingham risk scores were first published in 1976 [144] several other cohort studies have developed their own risk equations including PROCAM [11], SCORE [46] and QRISK [124]. These cohort studies differ significantly in terms of study population characteristics, risk predictors, and outcome [189].

A systematic review [189] identified 21 risk scores from 18 papers. Five were from Framingham [3, 8, 144, 314], three from the Munster group (PROCAM) [11, 12] and ARIC (Atherosclerosis Risk in Communities) [34, 200], two each from QRISK [124, 125, 123] and Reynolds [253, 254], and one each from the Scottish Heart Health Extended Cohort [318], Strong Heart Study [185], USA-PRC (People's Republic of China Collaborative Study of Cardiovascular Epidemiology) [320] and NHEFS (National Health and Nutrition Examination Survey NHANES I Epidemiological Follow-up Study) [93]. Some risk scores used multiple cohorts: SCORE [46] was derived from a pool of 12 European cohorts, and Progetto CUORE [233] from a pool of Italian cohorts. Twelve are from North America, eight are European, and one from China.

The use of routinely collected data from general practice patient records has been proposed for this research project and the access to CPRD (www.cprd.com) has been agreed for this purpose. The merit and appropriateness of using this type of routinely collected primary care data for the development of predictive models for cardiovascular disease has been validated by the work of Hippisley-Cox et al. [124, 125, 121, 120, 122] in developing QRISK and QRISK2. QRISK and then later the improved QRISK2 are cardiovascular risk prediction models (or risk calculators) for the general population. This approach has been further endorsed by the recommendation from National Institute of Clinical Excellence (NICE) that QRISK2, due its improved predictive power and better calibration with UK population, be used for predicting cardiovascular risk in UK general practice [217].

3.2 Cardiovascular Risk Scores for Type 2 Diabetes

People with T2DM have a two-fold increase in the risk of CVD [293, 317]. National policies for the management of both CVD and T2DM advocate the calculation of CVD risk in order to identify high-risk patients for targeted interventions [264, 295, 245, 66, 216]. Several multivariable risk-prediction models have been developed for the general, non-diabetic population that also account for diabetes, but only a few are specific to T2DM [303]. A systematic review by *Van Dieren et al.* [303] identified twelve studies where CVD risk scores were developed specifically for people with T2DM, and thirty-three were were developed for the general popu-

lation with diabetes as a risk factor in the model. The majority of these models were developed from a predominantly white population, and twelve were developed in Asian populations (India, China, Japan) [303] . Study sample sizes ranged between 698 and 1.5 million subjects.

Of the diabetes-specific models only two have been developed in patients with newly diagnosed T2DM , and they both use data from the United Kingdom Prospective Diabetes Study (UKPDS) [298]. These two models—one for Coronary Heart Disease (CHD) and the other for stroke—combine to form the UKPDS-RE [165, 284]. The other study populations composed of subjects with varying durations of diabetes. The majority of these models predicted 5-year risk and used and average of eight predictors. The most commonly used predictors were age, gender, duration of diabetes, HbA_{1c}, and smoking status.

Discriminative ability or *discrimination* is the ability of a risk score to differentiate between patients who did and did not experience an event during the study period. This measure can be quantified by calculating the area under receiver operating characteristic curve (AUC) [107] (equivalent to the *c* statistic for a binary outcome), in which a value of 0.5 represents random chance, while 1 represents perfect discrimination. Nine of the twelve studies reported the discrimination of their risk scores, with AUC values ranging from 0.68 to 0.85 [303]. In terms of AUC, values of 0.7 up to 0.8 indicate acceptable model discrimination [55, 131]. *Calibration* refers here to how closely the predicted x-year cardiovascular risk agreed with the observed x-year cardiovascular risk. Eight of these studies reported measures of calibration using Hosmer-Lemeshow goodness-of-fit statistic [131], p-values for which were all >0.05, indicating no significant lack of calibration [303]. Only half of these model internally validated their performance using resampling (e.g. split-sample, cross-validation, bootstrapping) techniques to correct for the overoptimism inherent when evaluating performance of models on the data in which they were developed.

The majority of models developed for the general population with diabetes as a risk factor predicted 10-year risk and used an average of eight predictors. The most commonly used predictors were age, gender, Systolic Blood Pressure (SBP), smoking status, lipid measurements and diabetes.

Twenty of the 33 studies reported discrimination, with AUC values ranging from 0.65 to 0.86 [303]. Twelve of these studies reported measures of calibration, all with Hosmer-Lemeshow

p-values > 0.05 [303]. Twelve of the models were internally validated using some form of resampling technique.

According to *Van Dieren et al.* [303], Thirty studies externally validated 14 different prediction models in patients with T2DM . Nine studies validated two versions of the UKPDS risk engine [165, 284], 10 studies validated three versions of the Framingham Prediction model [3, 314, 9, 8] and nine other prediction models were externally validated only once.

The UKPDS-RE CHD equations [284] have been externally validated in 10 separate studies [103, 139, 283, 322, 155, 237, 58, 304, 280, 302] which observed moderate discrimination and poor calibration, overestimating risk. The UKPDS-RE stroke equations [165] were validated in two separate studies [155, 58] with contrasting results, one reporting moderate discrimination and the other reporting good discrimination and good calibration. The largest of these studies from the UK used only a small sample (n=798) and from a single locality [283]. The largest international study had a larger but still relatively small sample size (n=7 502), using data collated from 20 countries [155]. The versions of the Framingham model by *D'Agostino et al.* [3] and by *Anderson et al.* [9, 8] were externally validated in a cohort with diabetes three times and the version by *Wilson et al.* [314] four times. Discriminative ability was moderate, with poor calibration. Of the risk scores that were only validated once, the Fremantle [58] score preformed the best with good calibration and discrimination.

Less than a third of the CVD risk scores identified in systematic review by *Chamnan et al.* [35] and *Van Dieren et al.* [303] were externally validated, with varying results. Both the diabetes-specific and general population risk scores reported good discriminative ability in the data in which they were developed (apparent and internal validation). However, their discriminative ability in new, previously unseen cohorts of patients (external validation) varied widely. The discrimination in models for the general population were generally moderate and the calibration mostly poor. As suggested by both *Chamnan et al.* [35] and *Van Dieren et al.* [303], this could be explained by differences in the incidence of CVD between patient with and without T2DM or by the fact that risk models developed in the general population do account for diabetes-specific risk factors. They argue that this could be overcome by using only diabetes-specific risk models for patients with diabetes. However, when the diabetes-specific risk models have been externally validated in new patients, the calibration was also poor with moderate to good calib-

ration. The more contemporary models such as the DCS [78], Fremmatle [58] and DARTS [72], appeared to have the best external validity, however these models were only validated once and therefore more external validation studies are required on different cohorts.

3.3 Genetic Programming in Bioinformatics

GP has long been applied to medicine, biology and bioinformatics. Early work by Handley [106] and Koza & Andre [169] used GP to make predictions about the behaviour and properties of biological systems, principally proteins.

Since then GP is has been widely used in biomedical data mining. Often the information that is of particular medical interest takes the form of very wide datasets, that is datasets with a relatively large number of columns, inputs or dependent variables [183]. Examples include chemical analysis using analytical techniques such as Fourier Transform Infrared Spectroscopy (FTIR), Gas Chromatography-Mass Spectrometry (GCMS), etc [114, 227, 140, 202, 290]; many more from Kell et al. (see below), Single Nuclear Polymorphisms (SNP)s where a typical data set may contain as many as 300,000 SNPs for 500–1,000 patients [19, 251, 272] and Affymetrix GeneChip microarray data [62, 80, 117, 126, 130, 174, 187, 190, 323].

Kell and his colleagues in Aberystwyth, and later Manchester, have had great success in applying GP widely in bioinformatics [5, 59, 79, 98, 97, 94, 94, 141, 153, 152, 154, 151, 273, 301, 316]. Another very active group is that of Moore and his colleagues in Vanderbilt , and later Dartmouth, [211, 213, 256, 257].

3.4 Genetic Programming in Prognostic Reseach

Albeit to a lesser extent than the biomedical domain, GP has been applied to specifically to medical diagnosis and prognosis [22, 192, 25, 193, 156]. However with the notable exceptions of Biesheuvel et al. [22], GP has been applied to data from clinical tests where there are a large number of readings but few samples, similar to biomedical data mentioned previously. This is contrary to the dataset proposed for this research project, which will come from general practice

records, where there are a relatively large number of samples and relatively fewer readings. This is also preferable in prognostic modelling, as a larger number of samples will increase the quality of the model, as too do fewer readings or inputs. A systematic literature search over PubMed, the largest database of biomedical publications, did not identify any applications of GP for medical diagnosis/prognosis using routinely collected primary or secondary care data, for which the data would consist of numerous samples in contrast to other types of biomedical data.

Biesheuvel et al. [22] compare GP and multivariate logistic regression in the development of a diagnostic prediction model using empirical data from a prospective diagnostic study among 398 patients in secondary care upon diagnosis of Pulmonary Embolism (PE) . Results report that the AUC of the GP model was significantly larger (0.73; 95%CI: 0.64–0.82) than that of the logistic regression model (0.68; 0.59–0.77), with comparable calibration or model-fit. The significantly larger AUC value suggests that GP may be better at discriminating those who experience the event of interest (for example, PE) versus those who do not. However where AUC values have been reported in the literature reviewed, most tend to be around 0.6–0.7.

However it is important to stress that improved discrimination of GP in diagnosis of PE reported by Biesheuvel et al. [22] is only a specific case. The 'no free lunch' theorem suggests that no single method will outperform all others on all cases [315]. Therefore it is important to further evaluate the performance of GP for medical diagnosis and prognosis.

3.5 Genetic Programming for Survival Analysis

The literature reviewed reports an enormous number of applications where GP has been successfully used [239]. The literature suggests that the specific GP task required by the project is that of symbolic regression [17, 167, 239]. Symbolic regression attempts to find a function that fits the given data points without making any assumptions about the structure of that function. Since GP makes no such assumption, it is well suited to this sort of discovery task [239]. Symbolic regression was one of the earliest applications of GP [167], and continues to be widely studied [30, 101, 149, 186].

Symbolic regression and GP appears to be well suited to classification and regression, and as discussed in the previous section, it has been successfully applied in clinical prediction modelling as an alternative for logistic regression where the data is uncensored [22, 136]. However, there is a surprising paucity in the literature regarding the application of GP to survival analysis and censored longitudinal data. There are lots of applications of GP for failure-time, reliability analysis, time-series, and other temporal applications in domains such as engineering and software development [327, 328, 188]. Surprisingly, there were no examples in the reviewed GP literature that involved the handling of censored longitudinal data, the experimental settings were such that every entity failed at least once and thus there was no censoring.

Expanding from the GP literature to the broader domain of AI and ML, there have been several studies, ANNs in particular, that have compared such novel non-linear statistical methods with their classic linear counterparts for survival outcomes [148, 138, 229, 248], however the results are mixed as to whether these non-linear methods offer improved performance. For example Schwarzer et al. [269] reviewed a substantial number of studies which have used ANNs in the diagnostic and prognostic classification in cancer, concluding that there is no evidence so far that application of ANNs represents real progress in the field of diagnosis and prognosis in oncology. Sargent [265] has also reviewed a number of these comparison studies showing that the majority have claimed equal performance but could not rule out the possibility of bias. GP , however, is a relatively recent technique that shows potential, which may improve the selection and transformation of predictors, and may lead to models with good predictive accuracy in new patients [22, 89, 136, 238, 297].

3.6 Summary Conclusions

In section 3.1 we introduce and review the large number of cardiovascular risk scores both for the general population that account for diabetes and for populations with T2DM . These models and they vary significantly in quality, development and validation methodologies, and in study population characteristics, risk predictors, and outcome. The degree to which these models have been assessed for generalisability in new cohorts also varies. The more contemporary models seemed to have the best external validity, however it is these models that have

only been validated once. Therefore more external validation studies are required to assess the performance of these models in different cohorts of subjects with T2DM . The moderate performance of most prediction model suggests that it is difficult to predict CVD in patients with T2DM [303].

In section 2.9 we introduce existing AI and machine learning techniques for survival analysis, which to date have been predominately ANN. We discuss how previous studies comparing the performance of these non-linear techniques to classic linear statistical approaches, have produced mixed results and that it remains unclear whether they offer any improved performance. We also discuss how GP is relatively recent approach with a lot of potential and that has been successfully applied for the purposes of symbolic regression in what is a well researched area.

In section 3.3 we discuss howGP has been applied with some success to the biomedical & health domain, often through the use of symbolic regression. However, the bulk of this research has been in biological domains such as modelling genomic data where there a disproportionally large number of columns which benefit significantly from the feature selection inherent in the evolutionary search process. To a lesser extent, GP has been successfully applied for prognostic research. However, with a small number of notable exceptions ,this has utilised biological data for prognosis rather than routinely collected longitudinal patient data.

In section 3.4 we discuss the results of out review of the research into the application of GP for prognostic research. The review did not reveal any research into using GP with routinely collected longitudinal health care data, which in contrast to biomedical data would have a relatively large number of rows and fewer columns. There was a single example by *Biesheuvel et al.* [22] where longitudinal patient data was used of prognosis, however it was using a small prospective cohort specifically designed for prognostic research rather than routinely collected data. The 'no free lunch' theorem [315] suggests that it is important to further evaluate the performance of GP for medical diagnosis and prognosis.

From the literature there appears to be only single example where GP has been used for prognosis in longitudinal patient-level data, however this example applied GP analogous to binary logistic regression where the outcome is binary. As discussed in section 2.8.2 this would discard valuable information and sacrifice statistical power.

There appears to have been a number of studies that have used machine learning, predominantly ANNs, for prognostic research where the outcome was time until event (i.e. a survival analysis). Whilst some of the techniques have been evolutionary in nature, GP was not considered. Despite the success GP in other areas of the biomedical and health domain, the utility of GP for clinical prediction modelling in the presence of censored data remain unknown.

In the next chapter we give an overview of the essential common themes in the diverse field of GP. We introduce a GP framework which later chapters will build upon, offering high-level overviews of the different elements, whilst also formally defining the specific methodological elements that will be implemented to form the developed GP approach for survival analysis.

Chapter 4

Genetic Programming

This chapter is give an overview of the essential common themes in the diverse field of GP. We introduce a GP framework which later chapters will build upon, offering high-level overviews of the different elements, whilst also formally defining the specific methodological elements that will be implemented to form the developed GP approach for survival analysis. After giving a high-level overview of a basic GP run, we introduce the concepts of solution representation and search spaces in GP, as well as approaches to define valid regions in these search spaces. Following on from this, we introduce ideas behind fitness and its role in guiding the GP search, before defining fitness functions used in this work. Then we introduce the role of GP search operators and define them. On that basis, we discuss different search strategies and define the search heuristics that will be used to implement them. Next we discuss the parameters that control the GP run and offer some guidance on suitable values for these. After which we discuss uncontrolled intron growth and the phenomenon that is bloat. Finally we discuss the considerations to be taken into account when implementing GP and define the specific implementation of GP used in this work, discussing the rationale behind certain choices.

4.1 Introduction

In GP we evolve a population of *computer programs*. That is, generation by generation, GP *stochastically* transforms populations of programs into new, hopefully better, programs (figure 4.1) [239]. For the context of this thesis a computer program is a symbolic regression equation, which a mathematical formula representing a clinical prediction model. Just like biological evolution, GP is a random (or stochastic) process and can never guarantee results.

However it is this randomness that enables GP to overcome the some of the pitfalls, i.e. local optima, that limit deterministic methods.

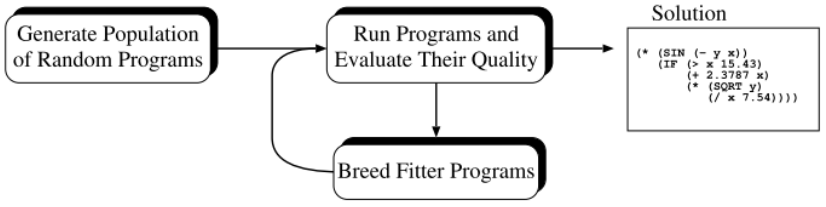


Figure 4.1: The basic control flow for GP, where survival of the fittest is used to find solutions (source: Poli at al., 2008 [239]).

Taken from Poli at al., 2008 [239], the basic step in a GP system are shown in Algorithm 4.1. GP finds out how well a model performs by applying it, and then comparing its behaviour to some ideal (line 3). For the purposes of this thesis we are interested in in how well a model predicts the risk of some clinical event . This comparison is quantified to give a numeric value called *fitness*. Those models that do well are chosen to breed (line 4) and produce new models for the next generation (line 5). Genetic variation operations are used to create (i.e. evolve) new programs from existing ones.

Algorithm 4.1 Genetic Programming

1: Randomly create an initial population of models from the available primitives (more on this in Section 4.2.1).
2: **repeat**
3: Execute each model and ascertain its fitness (Section 4.2.2).
4: Select one or two model(s) from the population with a probability based on fitness to participate in genetic operations (Section 4.3.4).
5: Create new individual model(s) by applying genetic operations with specified probabilities (Sections 4.3.2 - 4.3.3).
6: **until** an acceptable solution is found or some other stopping condition is met (e.g., a maximum number of generations is reached).
7: **return** the best-so-far individual .

4.2 Search Spaces & Fitness in Genetic Programming

This section provides an introduction to genotypic representations of individual solutions that, that along with GP search operators, define of the search spaces that GP will explore. We also introduce measures of quality, or fitness, of solutions in the search space. We discuss fitness in the context of survival analysis before formally defining the fitness measures that will be used in this work.

4.2.1 Search Spaces in Genetic Programming

In its simplest form, *representation* is the data structure used to define the the solution, i.e. an individual. However, we can also think of representation as the approach we take constructing, modifying, and presenting the individual for fitness assessment. Since early experiments into the automatic generation of executable structures [85] a variety of different representations have been explored including binary string machine code [91], finite state automata [86], generative grammatical encodings [309, 230, 65] and the dominant tree-based form [166]. Numerous alternative representations have also been proposed, including graph [291], strongly-typed [210], linear [27], linear-tree [145], and linear- graph [146].

Unsurprisingly given our knowledge of the No Free Lunch theorem [315], with so many different representation schemes the GP literature suggests that identifying appropriate representation schemes for certain type of problems remains an open issue, and as such is an active area of GP research [231]. There have been novel approaches proposed to address this issue, exploring the idea that GP can potentially evolve aspects of its own representation, such as Langdon's research on evolving data structures [180], Spector's investigations on *autocontructive evolution* [281], which co-evolve the search operators in addition to the individuals [231]. Another more recent approach is that of *GP Hyperheuristics* that look to explore search the space of all alternative algorithms and representations. In their review of open issues in GP O'Neil et al. [231] comment that Hyperheuristics are demonstrating early potential to outperform classic GP, and in an exciting twist, theoretical analysis suggests that a Free Lunch may be possible through their adoption [240, 241]. Whilst there is a diverse landscape of different representation schemes, the *tree-based* representation scheme popularised by Koza (1992) [167] remains

the dominant form and is the focus of this work.

Consider the tree in figure 2.8, containing the symbolic expression $max(x + x, x + 3y)$. This is the *parse tree* of a simple program which performs this mathematical operation, consisting of nodes that are elements of a *terminal set* and a *functional set* described in table 4.1. The leaves of the tree are input (predictor) variables or constants forming the terminal set, while the internal nodes are arithmetic operators that form the functional set. Together the terminal and functional sets form the *primitive* set of the GP system.

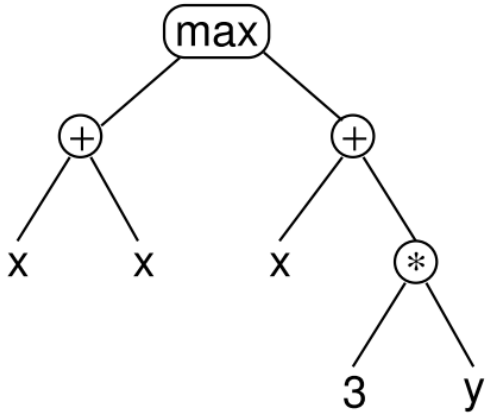


Figure 4.2: GP syntax tree representing $max(x + x, x + 3y)$ (source: Poli at al., 2008 [239]).

The functional set can be further subdivided into *binary nodes*, that take two inputs to produce an output, and *unary nodes* that take a single output. The primitive set represents the building blocks (or genetic material) that will represent the potential solutions to the problem, the individuals. Over successive generations the search operators (discussed in section 4.3) and fitness function will be applied to these building blocks to evolve populations of individuals towards a suitable, hopefully optimal, solution.

By defining the set of primitives, i.e., the functions, input variables, constants, along with the set of search operators (discussed in section 4.3), we define the *search space* that GP will explore. This consists of all possible solutions that can be constructed by the set of search operators in combining the primitives.

Table 4.1: Example of primitives in GP terminal and functional sets.

Set Type	Kind of Primitive	Example(s)
Function Set	Arithmetic	$+, *, /$
Function Set	Mathematical	$\sin, \cos, \exp$
Function Set	Boolean	AND, OR, NOT
Function Set	Conditional	$IF - THEN - ELSE$
Function Set	Looping	$FOR, REPEAT$
Terminal Set	Variables	x, y
Terminal Set	Constants	$3, 0.45$
Terminal Set	0-arity functions	$rand, go - left$

Constraing Genetic Programming Search Spaces

In order to work correctly functional primitives in GP require an important property called *closure* [167]. Closure is required because subtree crossover (described in section 4.3.3 can mix and join nodes arbitrarily. Koza [167] describes closure as being satisfied when each function is able to accept the output of any other function or terminal as in an input and remain syntactically correct. In traditional GAs, closure is not required as the chromosome is not treated as an executable programme [178].

Closure is the mechanism by which valid GP search spaces are defined. In GP there are two main approaches to ensure closure and define valid regions in the search space, often referred to as constraint handling in the EA literature. The first approach is to exclude invalid regions of the search space. Typically this achieved by constraining the GP system structurally or though some *type system*. Examples that fit into this category include *simple structural enforcement*, *strongly-typed GP*, and *grammar-based constraints*. This class of approaches result in sharp boundaries or *margins* in the valid regions of the search space.

The second and most common approach is to suppress invalid regions though fitness *penalty functions*, that is invalid individuals are given a significantly reduced fitness value. This kind of approaches leads to valid regions with softer margins. However, due to some the chal-

lenges associated with penalty functions [252], different approaches to automatically define good penalty factors and the development of alternative methods to handle constraints is an active area of research in GP [40, 41, 42, 49, 207, 279, 171, 252]. Several variations of penalty functions have been proposed in the EA literature including static, dynamic, annealing, adaptive, co-evolutionary, and death penalties [40, 207]. Alternative methods that have been proposed include repair functions that attempt to 'repair' infeasible solutions making them feasible, functions that separate objectives and constraints in MO optimisation settings, and hybrid approaches that combine other techniques such as Lagrangian multipliers or fuzzy logic.

Both these high-level approaches have their relative strengths and weaknesses, and there are some important considerations. Excluding invalid regions leads to a smaller search space. It does necessarily apply that a smaller search spaces mean more tractable problems as sometimes important intermediate solutions are excluded that assist the search in finding (near-) optimal solutions. However, if there is domain knowledge that strongly suggests a particular syntactic constraint on the solution, ignoring it will make is much harder to find a suitable solution.

All the experiments in this thesis utilise untyped GP and closure is partially satisfied by considering inputs coded as numeric variables. In our GP system implementation, invalid regions include invalid mathematical operations (e.g. dividing by 0) and solutions that do not represent time (e.g. do not include the time indicator variable, j in some fashion). Where genetic operations result in invalid solutions, then fitness penalised by setting the fitness value to ∞ (it the context of a minimisation problem). This approach, often referred to as the *death penalty*, effectively rejects individuals and is arguably the easiest way to constrain the search space and is computationally very cheap. Because invalid solutions have their fitness set to an extreme value, their fitness does not need to be calculated and no further calculations are required to quantify the degree invalidity or infeasibility the solutions. A key drawback of this is approach is that does not exploit any information from the infeasible solutions that may be generated by GP to inform the search.

There have been a number of publications in the active research area of constraint handling that suggest that the death penalty in not a good approach. For example, Coit & Smith [43] compared the death penalty against an adaptive penalty for a problem with a highly-constrained search space and found that the adaptive penalty approach was superior. Michalewicz [208,

209, 206] has shown that death penalty is inferior to penalties that are defined in terms of distance from the feasible region.

However, as small comparative study by Coello Coello [40] suggests that even the use of the death penalty may be sufficient in some applications, if nothing is known about the problem. However, its important to note that the same authors report that most comparative studies on constraint handling techniques in the literature are inconclusive and cite the the "No Free-Lunch Theorem" of Wolpert & Macready [315], suggesting that its expected that the best constraint handling techniques for a certain type of problem will tend to exploit specific domain knowledge.

Despite the limitations of the death penalty, the justification for its use as the constraint handling approach for this work was its computationally efficiency and ease of implementation combined with the fact that little is known about the application of GP to survival analysis. As discussed in later chapters, the computational expense of applying the developed GP approach to large clinical datasets was a recurrent challenge, and as pointed out by researches in the field, in some applications, the problem of finding a feasible solution might be itself NP-hard problem [279].

4.2.2 Genetic Programming Ftiness Functions

Fitness is some numeric measure of quality of each solution within the search space. The *fitness function* is the interface between the solution space, and the selection operator, that guides the GP search to regions or elements of the search space that consist of high quality solutions that solve (or approximately solve) the problem at hand. In GP, high quality solutions are 1) accurate, 2) parsimonious, 3) interpretable and 4) generalisable.

Accuracy

In GP, solution *accuracy* is highly problem dependant. For example a measure of accuracy for a classification problem would be very different from a metric that would quantify accuracy in a regression problem. In symbolic regression, the focus of this thesis, accuracy is typically measured by well-known error measures, which are the differences between observed $\hat{y}$ and

expected values y. In GP, commonly used error measures include mean-absolute error (MAE, Equation 4.1), sum-square error (SSE, Equation 4.2), mean-square error (MSE, Equation 4.3), root-mean-square error (RMSE, Equation 4.4), and scaled-mean-square error measure (SMSE, Equation 4.5), where scaling constants a and b are calculated via a least-squares fit. However, a limitation of these measures is that they cannot be used to compare models across different datasets because they are dependent on scale. Although slightly more computationally expensive to calculate, R^2, the proportion of variance explained by a model, is scale-free. However, this is constrained by the assumption of linearity in the model, which would be violated if the data significantly non-linear.

$$mae(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4.1}$$

$$sse(\hat{y}, y) = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4.2}$$

$$mse(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4.3}$$

$$rmse(\hat{y}, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4.4}$$

$$smse(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - (a + by_i))^2 \tag{4.5}$$

Parsimony

However, solution *parsimony* is largely problem independent. This can be measured in a number of ways for abstract genotypes. Typically parsimony measure appropriate to tree-based GP include tree size and tree depth. Tree *size* is merely the number of nodes in a (sub) tree. Tree *depth* is the minimal number of nodes that must be traversed to get from the root node of the tree to the selected node. Whilst being intuitive measure complexity, these and many other metric suffer from their coarse granularity. That is, significantly different trees may end up with same complexity value using these metrics. To address this issue, Keijzer & Foster (2007) [150] introduced a measure of tree *visitation length*, vl, defined in equation 4.6, where $c(t)$ denotes the number of children of the root of tree t and t_i denotes the ith subtree of t [150].

$$vl(t) := s(t) + \sum_{i=1}^{c(t)} vl(t_i) \tag{4.6}$$

For trees t of fixed size $s(t)$, $vl(t)$ will give smaller values for balanced trees, and larger values to unbalanced trees. It can be used to steer the GP search towards smaller well balanced solutions [83].

Interpretability

Solution *interpretability* is difficult to define explicitly and its utility is highly dependent on the application domain. For example, with solutions to sales and marketing problems, the application domain rarely cares how the solutions works, just that it works to increase profitability. In our case, the medical domain, interpretability is vital because it enables clinical validation. A clinical prediction models that is an excellent predictor of some important clinical outcome (like cardiovascular disease), is highly unlikely to be adopted or given any credibility if domain experts cannot understand how and why its such a good predictor. Whilst we cannot mathematically quantify it, solution interpretability is certainly correlated with solution parsimony.

Generalisabilty

The *generalisabilty* of a solution relates to its performance in a dataset other than one in which it was trained. The phenomenon of *overfitting* occurs when models learn not only the true pattern in the data, that may generalise to other data with similar characteristics, but also learn the pattern of the noise which is specific only to the training data. There are many approaches to evaluating including split sample approaches and resampling approaches, such as bootstrapping and cross-validation. Typically, more parsimonious solutions offer greater generalisability, although this isn't always the case.

Multi-objectve Fitness

In some cases it may advantageous to combine two or more different concepts that are on competition with each other into the fitness function. For example we could combine terms

solution complexity, and accuracy. Such a fitness function is referred to as a *multi-objective fitness function*.

State of the art GP systems are typically Multi-Objective Genetic Programming (MOGP) systems. In MOGP we optimise with respect to multiple competing fitness measures, $f_1, f_2, ..., f_n$, simultaneously. The complexity of solutions is one of the most important things to control during a GP. In some cases a significant increase in the complexity of individuals is observed, without a significant increaser in the fitness of the solutions. This leads to a phenomenon referred to as *bloat* (discussed in section 4.6), which can be disastrous for the GP run. Therefore MOGP provides an important mechanism for controlling complexity, whist simultaneously optimising other objectives such as the fitness of solutions

Many different approaches have been proposed to achieve multi-objective optimisation in GP. The most simple approach is to combine these fitness vectors in a single *aggregate scalar fitness function*, such as a weighted sum. This method has been used frequently in GP to control bloat. By combining program fitness and solution size to form a parsimonious fitness function one can evolve solutions that satisfy both objectives [167, 326, 325, 324]. There are many other examples in the literature of linear and semi-linear aggregations of fitness with other objective vectors [181, 18, 168]. However, these different fitness vectors are typically on different scales and thus estimating appropriate weights can be non-trivial. Furthermore, Pareto-optimal (discussed in subsection 4.3.4) solutions may be become unreachable. Typically in modern MOGP systems the objectives are kept separate, with fitness defined as a vector of real numbers. These fitness vectors then utilised by multi-objective selection operators, that operate on the notion of *Pareto dominance*, which is described in subsection 4.3.4.

Fitness in the Context of Survival Analysis

However, none of the well studied fitness (error) measures discussed thus far are suitable fitness measures for symbolic regression in the presence of censored survival data. This because unlike simple linear regression, where these error measures would be appropriate, there is no single continuous outcome y, with which to compare its distance from a models estimate $\hat{y}$. Rather, in survival problems we have a two-part outcome, with a continuous time until event value and

a dichotomous event indicator value.

As discussed in section 3.5, there is a surprising paucity in the literature regarding the application of GP of survival analysis and censored longitudinal data. Expanding from the GP literature to the broader domain of AI and ML, there have been some studies, ANNs in particular, that have evaluated the performance of non-linear methods for survival outcomes [148, 138, 229, 248, 255, 23, 158, 329, 172, 271]. These applications often utilise either Martingale residuals [255, 158, 329] or MLE [271] in some manner to develop goodness-of-fit (i.e fitness) measures in the presence of censored survival data.

The Martingale residual may be thought of as the difference between the observed and expected number of events for the ith individual. In the context of survival analysis, specifically where an individual can experience an event only once or not at all, Martingale residuals assess the relative magnitude of individuals time-to-event in comparison to what it predicted by a fitted model [277]. Martingale residuals have a mean of 0 across subjects and range between $-\infty$ and 1. Positive residuals indicate that the event occurred and that it occurred earlier that predicted - that the model "overpredicts". Negative residuals indicate that either the event did not occur (i.e. event time was censored); or the event occurred later than predicted - that the model "underpredicts". Equation 4.7 defines Martingale residuals for the i-th individual as:

$$\hat{M}_i = \delta_i - \hat{\Lambda}(t_i) \tag{4.7}$$

where δ_i is the number of events for the i-th subject between time 0 and t_i, and $\hat{\Lambda}(t_i)$ is the expected numbers based on the fitted model.

However there are potential issues or limitations associated with Martingale residuals. As a consequence of their definition (equation 4.7) Martingale residuals have a maximum of 1, are skewed towards negative numbers, and individuals that experience the event have, on average, larger martingale residuals than those with censored event times, which makes it difficult to use Martingale residuals to identify poorly predicted cases [277].

4.2.3 Devloping Ftiness Functions for Survival Analysis

Rather than a Martingale-based approach, this section proposes a MLE-based fitness function—that utilises LL as a goodness-of-fit measure—for survival analysis in censored longitudinal data [271]. Where LL (discussed in section 2.8.2) is used to calculate the distance between the natural log of the predicted probability p for the event to the actual observed outcome y.

In subsection 2.8.2 we introduced survival analysis, and the fundamental quantities used to assess the risk of event occurrence, and probability of being event-free, at a given time point as the *hazard* and *survival* functions, respectively.

In order to develop a MLE-based GP fitness function for survival data we take advantage of the fact that the hazard function corresponds to a conditional probability in the discrete time domain. Below, Equation 4.8 defines the *discrete-time hazard* function, denoted by $h(t_{ij})$, which is the conditional probability the individual i will experience the event in time period j, given that they did not experience it in any earlier time period and the their particular values of the set of covariates, X, in that time period [277].

$$\hat{h}(t_{ij}, X) = P[T_i = j | T_i \geq j, X] \tag{4.8}$$

This is in contrast to hazard in the continuous-time domain, which represents a rate, and as such can take values greater than one. The corresponding *discrete-time survial* function, denoted $\hat{S}(t_{ij}, X)$, is defined below in equation 4.9.

$$\hat{S}(t_{ij}, X) = \prod_{k=1}^{j} \hat{h}(t_{ik}, X) \tag{4.9}$$

It follows that the original survival analysis problem can be cast into a classification problem that requires the estimation of a conditional probability. However, to address the problem of censoring the data needs to pre-processed into the *counting process format*, where there are multiple rows per subject, one for each observed discrete-time interval. An example of survival data in this format is given in table 4.2, where X is the set of P covariates, and x_1 represents a time-independent variable (e.g. gender), and x_P a time-varying covariate (e.g. cholesterol). Note this advantageous feature of the counting time format, that it can inherently represent a combination of time-varying and static covariates.

Table 4.2: Example of survival data in the counting process format

PATID (i)	Time (j)	Event	x_1	...	x_P
01	1	0	1	...	0
01	2	0	1	...	1
02	1	0	0	...	1
02	2	0	0	...	0
02	3	1	0	...	1

Now that we have the data in this format we can reformulate $h(t_{ij})$ as the conditional probability $h(t_{ij}) = P(EVENT|X')$, where X' is a vector, consisting of the original vector of covariates (or features), X, plus an additional time period indicator, j.

Now we can estimate $P(EVENT|X')$ using the likelihood and prior ratios with a logistic link function. Below, equations 4.10 to 4.10 shows the derivation of the logistic link function.

$$h(t_{ij}) = \frac{1}{1 + e^{-\varepsilon}} \tag{4.10}$$

In the case where ε is a linear combination of covariates X' (including time indicator j), this represents a logistic regression model, which can be optimised using standard statistical techniques such as *Newton-Raphson method*. Below, Equation 4.11 defines ε_{lp} a linear predictor for discrete time survival analysis.

$$\varepsilon_{lp} = \{[\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + ... + \alpha_J D_{Jij}] + [\beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_P X_{Pij}]\} \tag{4.11}$$

In this definition D_{ij} is a 'dummy' time indicator (described in table 4.3), a dichotomy whose value indexes the time period j in the ith individual, P is the number of predictors (or covariates) and J is the number of observed time periods.

However, if we adopt a more complicated relationship for ε using a symbolic expression, we can model a non-linear relationship between hazard and covariates. It can be optimised by GP search operators, using the following likelihood function. Below, Equation 4.12 defines the

Table 4.3: Example of 'dummy' time indicators

PERIOD	D_1	D_2	...	D_{J-1}	D_J
1	1	0	0	0	0
2	0	1	0	0	0
...	0	0	...	0	0
$J-1$	0	0	0	1	0
J	0	0	0	0	1

Likelihood function for the discrete-time hazards model $Likelihood$.

$$Likelihood = \prod_{i=1}^{n} \prod_{j=1}^{J_i} h(t_{ij})^{EVENT_{ij}} (1 - h(t_{ij}))^{(1-EVENT_{ij})} \tag{4.12}$$

Here $EVENT_{ij}$ is a dichotomy representing the event indicator of the ith individual at the jth time interval. We define n is the number of subjects in training data and J_i as the number observed time periods (or terms) the ith individual contributes to the likelihood function.

To make optimisation through the GP search operators more computationally tractable, we take the logarithm of the likelihood to form a fitness function for survival analysis in censored data of the counting process format. Below, Equation 4.13 defines the *fitness function* ff_{surv}.

$$ff_{surv} = -\sum_{i=1}^{n} \sum_{j=1}^{J_i} EVENT_{ij} \log h(t_{ij}) + (1 - EVENT_{ij}) \log(1 - h(t_{ij})) \tag{4.13}$$

The fitness function expresses the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown population parameters.

There are similarities between this and the Martingale-based approaches in the sense that they both attempt to quantify some distance (or error) between predicted likelihood of events and observed event outcomes, inherently accounting for censoring. These two approaches differ in that Martingale approaches use event times and rates (e.g. the cumulative hazard function) whereas the MLE approach uses conditional probabilities in the discrete time domain.

This does give the Martingale approach the advantage of not requiring the data to be processed into discrete time segments, as is required for the MLE approach. However, the Maximum Likelihood-based method does not appear to suffer from the same biases as the Martingale-based approaches. This, and the fact that most logistic regression techniques use MLE to estimate the unknown parameters, were reasons that a MLE-based method was developed for this work in favour of a Martingale-based one.

4.3 Genetic Programming Search Operators

In GP the evolutionary process is implemented through the application of *search operators*, i.e., operators for population initialisation, selection, and variation. These operators define how the GP system will navigate through the set of possible genotypic solutions. GP departs significantly from other evolutionary algorithms in the way it implements variation operators, and the actual implementation of initialisation and variation operators is often specific to the GP representation scheme being used [239].

This work focuses on the well-known traditional approach to GP popularised by Koza (1998) [167], which use binary parses trees to represent individual solutions. In subsections 4.3.1-4.3.4 we introduce initialisation, mutation, recombination, and selection operators, appropriate for tree representations, that will be employed in our subsequent experiments. We also briefly review some of the other tree-based search operators, but a full review would be outside the scope of this work. For a detail review see excellent introductory texts from Poli et al., (2008) [239] and Banzhaf et al., (1998) [17].

4.3.1 Initialisation Operators

The first step in performing GP is the initialisation of a starting population. This involves the application of *intialisation operators* to create a variety of trees for subsequent evolution. Initialisation may also be used by certain variation operators and search heuristics to generate new subtrees, as a means of injecting some new genetic material into the population during the GP run which helps to preserve *Population diversity*.

Population diversity is valuable in property in GP. The trees developed can vary in term of their size, depth, width, and the number of leaf (i.e. terminal) nodes. Whilst the shape of the initial trees can be lost in a few generations, the diversity of the of the initial population is an important factor that can effect the quality of the solutions that the GP system can develop. The initial population provides (the majority of) the building blocks or genetic material that will be evolved (section 4.3) to (hopefully) find a high quality solution to the problem. Thus if the diversity of trees in the initial population is low or biased (i.e. they are very similar) then the GP system has less to work with, conversely if the initial population is diverse the the GP system and more to work with. Whilst the application of genetic variation operators, either thought mutation or adding newly initialised individuals, can introduce new, previously unseen genetic material (discussed in section 4.3), these are typically applied at a low rate and therefore the (genetic) diversity of the initial population is important.

Random Initialisation

Like most evolutionary algorithms, in GP an initial starting population is typically generated randomly. The elements of the primitive set (section 4.2.1), the GP building blocks, are combined randomly to produce a initial population of trees.

There are many different approaches to randomly generating this initial population of trees. In the *full* method (so named because it generates full trees, i.e. all leaves are at the same depth) nodes are taken at random from the function set until the maximum tree depth is reached (beyond that depth, only terminals can be chosen) [239]. Although the full methods generates trees that are all of the same depth, this doesn't necessarily mean that all the tress are of the same size (i.e. the total number of nodes) or shape. This is only the case when the function set contains only functions with the same arity. The *arity* of a function is the number of the arguments accepted by a function. However, even when the function set consists of mixed-arity functions, the sizes and shapes of the trees generated by the full method tend to be rather limited [239].

In contrast, the *grow* method produces irregular trees because the nodes are selected from the whole primitive (i.e. functions and terminals) set at random throughout the entire tree [17].

Once a branch has selected a terminal, that branch is ended, even if the maximum tree depth is reached. If the tree depth is reached by a given branch then a terminal node is selected. A key limitation of the full and grow initialisation operators is that neither provide a diverse range of sizes and shapes, and can lead to a uniform set of structures in the initial population because the routine is the same for all individuals [17].

To prevent this Koza [167] proposed a combination of these two methods, called the *ramped half-and-half* method, intended to enhance the population diversity from the outset. Half the initial population is constructed using full and half is constructed using grow. This is done using a range of depth limits (hence the term "ramped") to help ensure that we generate trees having a variety [239]. The ramped half-and-half method most commonly used initialisation operator in tree-based GP.

Below, Equation 4.14 defines the *random initialisation* operator, $init$. Here, $\mathbb{P} \in [0, 1]$ denotes the set of probabilities, i.e., real numbers in the interval between 0 and 1. The set of functions with arity equal to or greater than one is denoted by $F_{>0}$ [83].

$$
\begin{aligned}
&init(n, p_s, p_v) = I(0, n, p_s, p_v) \\
&I(i, n, p_s, p_v) = \begin{cases} f(\underbrace{I(i+1, n, p_s, p_v), \ldots, I(i+1, n, p_s, p_v)}_{arity(f)}) & \text{if } r_u(0,1) < p_s \text{ and } i < n. \\ r_{du}(V) & \text{if } r_u(0,1) < pv. \\ r_{du}(F_0) & \text{otherwise.} \end{cases} \\
&f = r_{du}(F_{>0})
\end{aligned}
\tag{4.14}
$$

Random initialisation creates trees of maximum depth n. A subtree is created with probability p_s at each recursive step. If no subtree is created, an input variable is created with probability p_v, else a constant is created. The full initialisation strategy can be realised by setting $p_s := 1$ and $p_v := |V|/(|V|+|F_0|)$, the ratio between the number of input variables and the number of terminals, where only full trees of depth n are created. By setting p_v as in the full strategy and $p_s := |F_{>0}|/|V|+|F|$, the ratio of the number of functions of arity equal to or greater than one and the number of all functions and input variables, the grow initialisation strategy can be realised.

Other Initialisation Operators

Whilst Koza's ramped half-and-half method is the most common, there are several other ways of constructing a population (pseudo) random individuals. As stated by Poli et al. (2008) [239], the shape of the initial trees can be lost within the first few generations and a good initial population can be crucial to the success of a GP run. The search space often consists of an infinite number of possible solutions and thus it is impossible to search them all in a uniform fashion. Therefore any approach to construct an initial population of solutions will be subject to bias [239].

For example, the ramped half-and-half method tends to produce bushy trees which may on average be better for some types of problems (such as parity problems), but may not be the best approach for other types of problems. This is demonstrated by the fact that the ramped half-and-half method is poor at finding solutions to the Sante Fe ant trail-following problem [176]. Another potential issue is that trees produced by the ramped half-and-half method may just be too small for some problems. Chellapilla (1997) [37] claims good results are achieved when the size of the initial trees was more tightly controlled.

Other methods have been proposed that sample trees uniformly based on *Alonso's bijective algorithm* [7, 132, 24, 173]. These more "uniform" initialisations on average tend to produce more asymmetric trees, in contrast the symmetric trees that are generally constructed by the ramped half-and-half method. Therefore, uniform sampling and other initialisation methods may serve as important alternative for certain problems.

Of course, the initial population need not be generated in a random fashion. If something is known about likely properties of the desired solution, whether produced from a previous GP run or perhaps constructed by the user, symbolic expressions having these properties can be used to *seed* the initial population [4, 129, 175, 179, 308]. However there are some considerations when seeding an initial population. Just as the shape of the initial trees can be lost within the first few generation, so can a few high-fitness solutions dominate the population in the first few generations, leading to rapid loss in genetic diversity. Poli et al. (2008) [239] suggest that diversity preserving techniques, such as MOGP [234, 270], demes [180], fitness sharing [96] and the use of multiple seed trees, might be good cures for the problems associated with the

use of a single seed.

4.3.2 Mutation Operators

Mutation only operates on one individual. Mutation operators were applied in early work in the evolution of programs [21, 51]. However, Koza, who is credited for popularising the field, did not use mutation in his seminal work in field [167]. This work has had a wide influence and mutation is often omitted. However, more recent comparisons of crossover and mutation suggest that including mutation can be advantageous [239]. Chellapilla (1997) [37] found that a combination of six mutation operators performed better than previously published GP work on four simple problems. Harries and Smith (1997) [113] also found that mutation-based hill climbers outperformed crossover-based GP systems on similar problems. In modern GP applications mutation is used widely. Whilst its true that you don't need to use mutation in GP, there doesn't appear to be any consensus in field on its relative merits. It has also been suggested that, when the problem is complex, the relative merits of variation operators are not only dependent on the problems but also on the actual implementation of the GP system [197].

Subtree Mutation

Below, Equation 4.15 defines the *subtree mutation* operator, mut_s. In this definition $v \in V$ denotes an input variable [83].

$$mut_s(f(t_1,\ldots,t_n),n,p,p_i,p_s,p_v) = \begin{cases} init(n,p_s,p_v) & \text{if } r_u(0,1) < p \text{ and } r_u(0,1) < p_i. \\ init(0,p_s,p_v) & \text{if } r_u(0,1) < p \text{ and } r_u(0,1) \geq p_i. \\ f(mut_s(t_1,n,p,p_i,p_s,p_v), & \\ \ldots,mut_s(t_n,n,p,p_i,p_s,p_v)) & \text{otherwise.} \end{cases}$$

$$mut_s(v,n,p,p_i,p_s,p_v) = \begin{cases} init(n,p_s,p_v) & \text{if } r_u(0,1) < p \text{ and } r_u(0,1) < p_i. \\ init(0,p_s,p_v) & \text{if } r_u(0,1) < p \text{ and } r_u(0,1) \geq p_i. \\ v & \text{otherwise.} \end{cases} \quad (4.15)$$

Subtree mutation is the most commonly used mutation operator is tree-based GP, where a mutation point is randomly selected and the subtree below this point is deleted and replaced with a randomly generated subtree or terminal node. The altered individual is then placed back into the population. The maximum size and shape of the newly generated subtree can be controlled by the parameters n, p_s, and p_v, with the same semantics as equations 4.14. The strength of the mutation operator is controlled by, p, the probability of replacing a subtree of a node with a randomly generated subtree at each recursion of the tree. The parameter p_i is the probability that a newly generated subtree will be a terminal, controlling the tendency of the operator to grow or shrink trees. Subtree mutation is described graphically in figure 4.3. The new randomly generated subtrees are typically produced according to the same initialisation scheme, with the same limitations (e.g. in terms of depth and/or size) as the initial population.

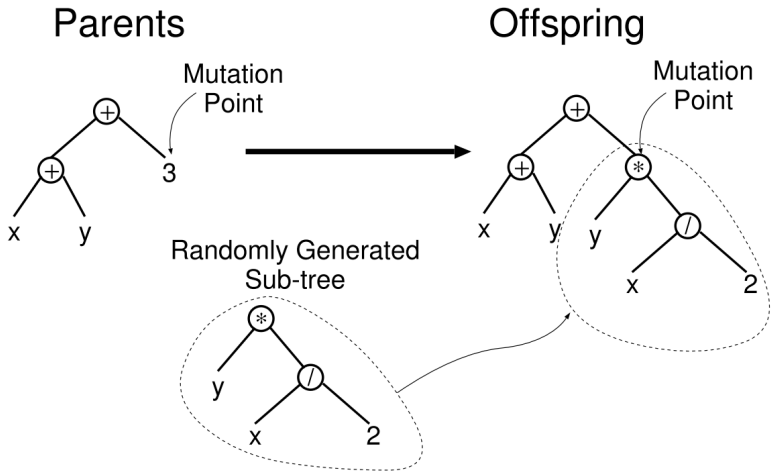


Figure 4.3: Example of subtree mutation (source: Poli at al., 2008 [239])

Point Mutation

Below, Equation 4.16 defines the *point mutation* operator mut_p. In this definition $v \in V$ denotes an input variable [83].

$$mut_p(f(t_1,\ldots,t_n),p) = \begin{cases} f'(mut_p(t_1,p),\ldots,mut_p(t_n,p)) & \text{if } r_u(0,1) < p. \\ f(mut_p(t_1,p),\ldots,mut_p(t_n,p)) & \text{otherwise.} \end{cases}$$

$$mut_p(v,p) = \begin{cases} r_{du}(V) & \text{if } r_u(0,1) < p. \\ v & \text{otherwise.} \end{cases}$$

$$f' = r_{du}(F_{arity(f)}) \tag{4.16}$$

Point mutation analogous to mutation in GA, where bits are flipped. In GP a node in a tree is selected at random and replaced with randomly selected node of the same arity and type. This way functions are replaced buy function and variables by variables. The probability to replace the current node when recursively traversing the tree, denoted p, controls the strength of the operator. It important to note that point mutation preserves the shape of the tree.

Constants at Random Mutation

Below, Equation 4.17 defines the *constantat at random mutation* operator, mut_c. In this definition $v \in V$ denotes an input variable. This operator assumes the presence of numeric constants $\mathbb{R} \subseteq F_0$ denotes an input variable [83].

$$mut_c(f(t_1,\ldots,t_n),p,\mu,\sigma) = \begin{cases} f + r_n(\mu,\sigma) & \text{if } r_u(0,1) < p \text{ and } f \in \mathbb{R}. \\ f(mut_c(t_1,p,\mu,\sigma), & \\ \ldots, mut_c(t_n,p,\mu,\sigma)) & \text{otherwise.} \end{cases}$$

$$mut_c(v,p,\mu,\sigma) = v \tag{4.17}$$

Here, $f \in \mathbb{R}$ denotes a real-valued function of arity, i.e., a constant. The constant at random mutation operator mutates constants by adding random noise drawn from a Gaussian distribution with mean μ and standard deviation σ. As before, the strength of the operator is controlled by, p, the probability of mutating a constant at each recursion. Each change to a constant is considered a separate mutation.

Other Mutation Operators

In contrast to GA where mutations is simply flipping a bit in a bit string, in GP there are many different types of mutation operators that can be applied to trees. Often multiple types of mutation are beneficially used simultaneously [170, 10]. Table 4.4 gives a brief overview of what types of mutation operators have been used with trees.

Table 4.4: Mutation operators applied in tree-based GP

Operator Name	Description of effect
Subtree	Subtree exchanged with random subtree [167, 160]
Size-fair subtree	Subtree exchanged with random subtree that is, on average, the same size [160]
Point	Single node exchanged with random node of the same class [203]
Hoist	New individual is generated from subtree [161]
Shrink	Subtree exchanged with a random terminal [10]
Expansion	Terminal exchanged with a random subtree
Permutation	Arguments (subtrees) of a node are randomly permuted [167, 201]
Constants at random	Add noise to constant, according to a Gaussian distribution [268]

4.3.3 Recombination Operators

GP *recombination operators* or *crossover operators* take genetic material form two selected 'parent' trees and swaps part of one parent with another, combing them to form one or two new 'offspring' trees. In GP the most common form of crossover is tree-based crossover, often referred to as *subtree crossover*. Subtree crossover is

described graphically in figure 4.4. The parents are shown on the left and the offspring on the right. More specifically:

- Using some selection scheme (described in section 4.3.4, choose two solutions from the current population to act as parents
- Randomly select a node in each parent to act a crossover point.
- Delete all branches (i.e. the subtree) below the crossover point in each parent (shaded areas in figure 4.4).
- Recombine (i.e. crossover) the remaining partial trees at their respective crossover points to form a new tree to act as an offspring .

Note, that it is possible to define a version of subtree crossover that can produce two offspring, but this is not commonly used [239].

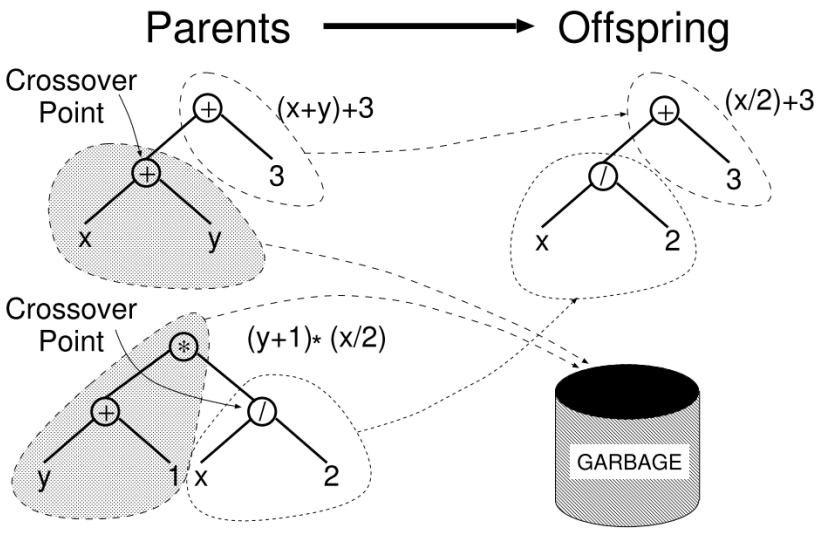


Figure 4.4: Example of subtree crossover. Note that the trees on the left are actually copies of the parents. So, their genetic material can freely be used without altering the original individuals (source: Poli at al., 2008 [239]).

Subtree Crossover

Below, Equation 4.18 defines the *subtree crossover* recombination operator, rec_s. [83].

$$\begin{aligned} rec_s(t,u) &= (t[\![i \mapsto u_{[j]}]\!], u[\![j \mapsto t_{[i]}]\!]) \\ i &= r_{du}(\{0,|t|\}) \\ j &= r_{du}(\{0,|t|\}) \end{aligned} \tag{4.18}$$

In this definition, $t[i]$ denotes the ith subtree of tree t, where $t[0] = t$. The notation $t[\![i \mapsto u]\!]$ denotes the tree t whose ith subtree has been replaced by tree u. The number of subtrees of a tree is written as $|t|$. The subtree crossover operator randomly selects a crossover point in each parent expression and swaps the corresponding subtrees, returning two offspring trees.

Other Recombination Operators

In this definition of subtree crossover the selection of subtrees, through the selection of a crossover point, is selected with uniform probability. Uniformly selected crossover points can lead to the crossover of limited genetic information (i.e. small subtrees) . To address this the selection of subtrees can be biased so that subtrees containing terminal are selected with a lower probability than other subtrees. Koza (1992) [167] proposed a widely used approach of choosing functions 90% and terminals 10% of the time to counter this problem.

During biological sexual reproduction, the genetic material from the parents appears in approximately the same place in the child. This is quite different from traditional tree-based GP crossover, which can move a subtree into a totally different position in the tree structure. *One-point* and *context preserving* are examples of *homologous* crossover operators, they preserve the position of the genetic material through the use of a *common* crossover point in the parent trees. There are many other types of homologous crossover operators that have also been proposed for tree-based GP [44, 173, 191,

198, 321]. The prominent crossover operators that are applicable to trees are detailed in table 4.5.

Table 4.5: Crossover operators applied in tree-based GP

Operator Name	Description of effect
Subtree	Exchange subtrees between individuals [167]
One-point	Exchange subtrees if coordinates match and subtrees have same shape [177, 242, 244]
Uniform	Exchange nodes randomly between individuals, with uniform probability [243]
Context preserving	Exchange subtrees if coordinates match [69]
Size-fair	Exchange subtrees between individuals that are, on average, the same size [173]

4.3.4 Selection Operators

As mentioned previously, in GP, the the evolutionary process is a search that is facilitated through search operators. As part of this search GP uses *selection operators* to choose m individuals, from a pool of n individuals, that will be subject to genetic variation (*parent selection*), or for transfer into the next generation (*survival selection*). As with most EAs, GP employs *fitness-based selection*, where individuals are selected as parents or survivors, either deterministically or probabilistically, based on fitness. That is better solutions are more likely to be selected as parents or survivors, than inferior solutions. The selection of individuals as parents or survivors is a trade-off between *exploitation* of high-quality individuals, and *exploration* of the search space thought through the selection of average-quality solutions (which may act as intermediate solutions in the search path later lead to individuals of even higher quality).

As well being either deterministic or probabilistic, search operators can be single-

objective or multi-objective. Furthermore, search operators can be classified along a spectrum between *non-elitist* and *elitist*, dependent on how they value exploration over exploitation. There are numerous selection mechanisms, which have been described many times in the EA literature. A full review of all these approaches is outside the scope of this thesis, Goldberg (1989) and Luke (2013) [96, 195] provide discussion on various selection mechanisms.

Single-Objective Selection

The most commonly employed *single-objective selection* operator in GP is *tournament selection* [239]. Figure 4.3 in section 4.4.1 gives a pseudocode implementation of the tournament selection operator used in this work. In tournament selection a number of individuals are selected at random from the current population to compete in the tournament. From the pool of competitors, the individual with the best fitness in selected as the winner of the tournament, and selected for genetic variation. If the genetic operator to be applied is crossover, two tournaments are used, one for each parent.

A key property of any selection mechanism *selection pressure*. A system with strong selection pressure very highly favours fitter individuals, whilst a system with low selection pressure isn't so discriminating [239]. Tournament selection automatically rescales fitness, keeping the selection pressure constant. This is because tournament only looks at the relative fitness, though ranking the competitors based on fitness, rather than how much fitter they are. In this way, an exceptionally fit individual cannot immediately swamp the future generation with its children, which would have disastrous consequences on the GP run due to a drastic reduction in the diversity of genetic material available to future generations. Conversely, tournament selection amplifies small differences in fitness, to prefer better solutions even if they are only marginally better than the other individuals in the tournament. In tournament selection, the *tournament size* parameter allow researchers to adjust the selection pressure, as such control the degree of elitism.

A disadvantage of tournament selection is the noise that is introduced thought the random selection of individuals for participation in the tournament. This mean that individuals with average fitness can have some chance of being selected and their offspring featuring in future generations. Despite this drawback, since tournament selection is easy to implement, offers some control of elitism, and offers automatic fitness rescaling, it is commonly used in GP.

Multi-Objective Selection

As discussed in subsection 4.2.2 the fitness function acts as an interface between the solution space and the selection operator, guiding GP though the search space towards high-quality solutions. In section 4.2.2 we also discussed that it is often highly desirable to employ a multi-facetted definition solution quality. Thus a fitness values can be a vector of different quality elements such as accuracy, parsimony, interpretability, and generalisability. Single-objective selection operators can only be used with fitness vectors when an aggregate scalar fitness value (such as a weighted sum) is used. However, as previously discussed identifying suitable weightings is non-trivial and often precludes this approach.

A alternate approach that is gaining popularity is to adopt a *multi-objective selection* operator than can handle multiple separate quality criteria directly. The most common forms of multi-objective selection are *lexicographic selection* and *Pareto selection*. Lexicographic selection uses a lexicographical ranking of fitness vectors. Using the example where accuracy and parsimony are the objectives; lexicographic selection ranks subjects based on fitness, and where there are individuals that are tied having the same fitness value, the individual with the lower complexity is assigned the higher ranking [195].

Pareto selection operators are based on the notion *Pareto dominance*. Given a set objectives and two candidate solutions, A and B. A is said to *Pareto dominate* B if A is at least as good as B in all objectives, and superior to B in at least one objective. You

could say that, A is at least as good everywhere and better in something. In Figure 4.5, an example taken from Poli et al. (2008) [239], individual A dominates (is better than) individual B along the y axis, but B dominates A along the x axis. Thus there is no simple ordering between then. This illustrates the notion of a *partial order*, where there is no longer a strict linear ordering of solutions. The individual marked '2', however dominates B on both axes and would thus be considered strictly better than B.

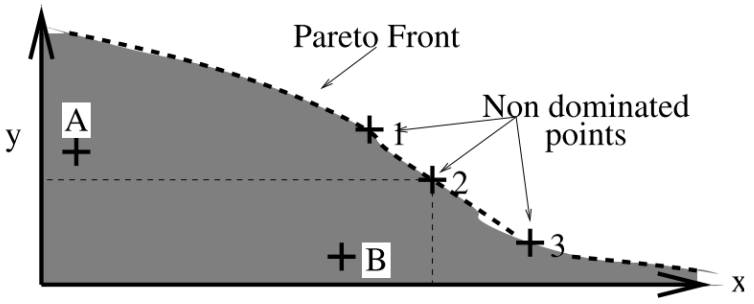


Figure 4.5: Two-dimensional example of Pareto optimality and the Pareto front, where the goal is to maximise along both the x and y axes. Solutions A and B do not dominate each other. However, solution B is dominated by solution 2. (source: Poli at al., 2008 [239]).

When A and B are identical in all objectives, or like in our example, if B is better in some things but A is better in other things, the solutions are said to be*non-dominated*. In this case the goal of the search operator is find the *Pareto font*, this is, the set of *all* non-dominated solutions in the search space.

In such a scenario there is said to be no *Pareto-optimal* set of solutions, but the selection operator needs to select some set of solutions. To address this we quantify how *close* solutions are to the Pareto front. One popular method is *Non-Dominated Sorting (NDS)* by Srinivas and Deb (1994) [282], which based on the notion of *Pareto front rank*. Luke (2013) [195] offer a nice introduction, describing NDS as "like peeling an onion". Fitness vectors in Pareto front get a rank of one. They are then removed, the front

recalculated, and the fitness vectors in this second Pareto font get a rank of two, and so on. NDS continues to iterate until every fitness vector has a Pareto rank.

Pareto selection operators selects fitness vectors from consecutive Pareto fronts until the specified number of m fitness vectors are selected. In some cases there may be more fitness vectors in a given Pareto front than are required. In this case the selection operator needs to adopt some strategy for choosing a subset of fitness vectors in a particular Pareto front. The most common strategies are based on *crowing distance* and *hyper volume contribution*, implemented by the NSGA-II [64] and SMS-EMOA [20] algorithms, respectively.

Pareto-based selection methods have a number of advantages. The main advantage (and the motivation behind their initial application in GP) is the control of bloat, where practitioners can manages the trade-offs between solution quality and solution complexity. Of course, dependent of specific problem, the simultaneous optimisation of other objectives can be advantageous for a GP run. Another advantage is that by basing selection criteria on multiple objects we can help ensure genetic diversity in the GP run. MOGP is increasing popular with many state of the art GP using Pareto selection. All of the experiments in this thesis use multi-objective selection based the Pareto dominance.

Lexicographic ordering and Pareto dominance are not the only ways to deal with multiple objectives without combining them into a single scalar fitness function. Alternate approaches have been proposed that include weaker forms of Pareto dominance and new methods based on priorities [267] and those which combine Pareto dominance and lexicographic ordering [196].

Population Diversity Preservation

Although much of the application multi-objective optimisation to GP has been in the area of bloat control, discussed in section 4.2.2, preserving diversity in the popula-

tion is another important consideration. As discussed in section 4.3.1, a rapid loss in genetic diversity can have disastrous consequences for a GP run. Without some mechanism to preserve diversity in a population (i.e. promote exploration), searches can easily converge to solutions that are too small to solve a problem. One such approach is use a multi-objective search with some measure of diversity as an objective, typically whilst simultaneously controlling for bloat (parsimony), and optimising fitness (accuracy) [258]. A particularly successful multi-objective approach by Schmidt and Lipson (2010) [266], the *age-layering technique*, uses the notion of the age (how long the genotype has been in the population) to ensure diversity. Using a multi-objective search operator, a dynamic variant of the age-layering technique is realised by adding solution age as an objective to be optimised. Many other approaches to diversity preservation have been proposed by the EA community including demes, random restarts, crowding, and fitness sharing.

4.4 Search Stratergy

Now that we have introduced the individual components such as primitives, initialisation, variation, fitness, and selection, we put them together using an overarching GP *search strategy*, or in machine learning terminology, a GP *search heuristic*. In GP search heuristics can be classified into two broad groups, generational and steady-state. With *generational* search heuristics, the entire populations is updated once per iteration. Whereas with *steady-state* heuristics, there are no generations *per se* and only a few individuals are updated per iteration.

Typically GP is implemented using a steady-state search with tournament selection. Modern GP systems typically use a multi-objective evolutionary algorithms (MOEA) as their search heuristic. The majority of today's best performing GP systems use steady-state algorithms with Pareto tournament selection. This popularity, whether for simple teaching examples or complex real-world systems, is that they are relatively

simple to implement and allow straight-forward parallelisation. Despite this, simple steady-state EAs with single-objective tournament selection are widespread [239].

This section describes the set of GP search strategies, group by steady-state and generation, that have been selected for this study. Search heuristics were selected from those available from the *RGP* package [84, 83] within the *R* statistical programming language [247], discussed in section 4.7. We implement the popular steady-state heuristic with single-objective tournament selection, SSOGP, that is implemented in many existing GP systems. To explore the utility of a generational approached have chosen to implement Generational Single-Objective Genetic Programming (GSOGP), and to explore the utility of multi-objective GP we have implemented Generational Multi-Objective Genetic Programming (GMOGP). Unfortunately there wasn't a multi-objective version of the steady-state search heuristic available in the RGP package at the time of writing. An overview of the important features and attributes of the search strategies selected for this work is given in Table 4.6.

Table 4.6: Overview of the important features and attributes of the GP search heuristics described in this work..

	SSOGP	GSOGP	GMOGP
Optimisation Criteria	Fitness	Fitness	Fitness, Complexity, Age
Selection Framework	Steady-state	Generational ($\mu+\lambda$)	Generational ($\mu+\lambda$)
Parent Selection	Uniform random	Uniform random or rank-based	Uniform random or NDS
Variation	$rec \rightarrow mut$	$rec \rightarrow mut$	$rec \rightarrow mut$
Survivor Selection	Rank-based	Rank-based	NDS
Diversity Preservation	-	-	Age-Fitness Pareto Optimisation (AFPO)

4.4.1 Steady-State Single-Objective Genetic Programming

In steady-state GP there are no fixed generational intervals. Instead the population is updated in a piecemeal fashion rather than all at one time, with a continuous flow of individuals meeting, mating and producing offspring. The idea is to iteratively breed a small number of offspring, access their fitness, and, if the they have superior fitness, these offspring replace existing individuals in the same population. The method is simple to implement and has some efficiency benefits together with benefits from parallelisation.

"The approach is called steady-state because the genetic operators are applied asynchronously and there is no centralised mechanism for explicit generations. Nevertheless it is customary in presenting results with steady-state GP to talk about generations. In fact steady-state generations are intervals during training which can be said to correspond to generations in a generational GP algorithm. These intervals are often when fitness is evaluated for the same number of individuals as the population size" [17]

This heuristic has been implemented as a minimal example of the most common and easy to implement GP search strategy. It is intended to act as baseline for comparison with more complex single- and multi-objective generational approaches in the development of clinical prediction model for censored data.

Algorithm Structure

Here we describe simple single objective steady-state GP as it is implemented in RGP, which is loosely based on Koza's original work on GP [167]. In the first step we initialise population $pop(0)$ with μ random individuals, before entering the main evolutionary loop. Next we randomly choose a variation operator, according to parameter p_{rec}. Note that the choice between recombination and mutation in mutually exclusive, that either variation operator is applied, but not both. If the variation operator is recombination, two parents are selected via two independent tournaments, each of size $s_{tournament}$, as

detailed in the next subsection. If the variation operator is mutation, then one parent is selected via a single tournament. Next, a single child is created by applying the variation operator to the parent(s). Then, an individual from the population is selected for replacement by the new offspring, via a single (negative) tournament, again, of size $s_{tournament}$. Finally, the newly created child is inserted into the population in place of the selected individual. This evolutionary process is repeated until the termination criteria is fulfilled. Pseudocode for this simple steady-state search heuristic is given in Algorithm 4.2.

Algorithm 4.2 Pseudocode implementation of the SSOGP search heuristic. [83]

```
1:  pop ← createIndividuals (number = μ)

2:  while termination criterion not met do
3:      if randomUniformNumber() ≤ p_rec then     ▷ uniform random numbers [0, 1]
4:          mother ← tournament(pop, s_tournament)
5:          father ← tournament(pop, s_tournament)
6:          child ← rec (mother, father)
7:      else
8:          parent ← tournament(pop, s_tournament)
9:          child ← mut (parent)
10:     end if

11:     replaced ← negativeTournament(pop, s_tournament)
12:     pop[replaced] ← child
13: end while

14: return pop
```

Selection Stratergy

Tournament selection is implemented in RGP as follows. First an individual is randomly selected, with a uniform probability and without replacement, from the population. This starting individual is designated as the *bestIndividual* and its fitness the *best fitness*. Then iterate the following steps until the number of individual selected equals $s_{tournament}$. We randomly select another individual from the population, in the same manner as before, to act as a *competitor*. If the competitor has a better fitness than the best individual, then it becomes the *bestIndividual* and its fitness becomes the *bestFitness*, else nothing. We then iterate through the for loop until we have evaluated $s_{tournament}$ individual. Finally, when the loop is finished, we return the *bestIndividual*. Negative tournament selection is very much the same process, only that we return individual with the worst fitness (rather than the best). Note that this is very simple implementation of single-objective tournament selection, in practice there are much more complicated variants and extensions. Pseudocode for this tournament and negative tournament selection is given in Algorithm 4.3.

Diversity Preservation

In the simple single-objective steady-state search heuristic implemented in RGP there isn't any mechanism to preserve genetic diversity in the population. However, it could be extended with minimal effort using external measures such as random restarts, fitness, sharing and crowding. Such extensions have not been implemented in this work for the sake of simplicity.

Parameters

Table 4.7 details the most important parameters appropriate to the RGP implementation of SSOGP. Selection of suitable stating parameter was discussed in section 4.5, but the

Algorithm 4.3 Pseudocode implementation of tourament selection [83]

```
1: function TOURNAMENT(pop, s_tournament)
2:     bestIndividual ← sampleWithoutReplacement(pop, number = 1)
3:     bestFitness ← ∞

4:     for i in 1 : s_tournament do
5:         competitor ← sampleWithoutReplacement(pop, number = 1)
6:         if fit(competitor) < bestFitness then
7:             bestFitness ← fit(competitor)
8:             bestIndividual ← competitor
9:         end if
10:    end for

11:    return bestIndividual
12: end function

13: function NEGATIVETOURNAMENT(pop, s_tournament)
14:    worstIndividual ← sampleWithoutReplacement(pop, number = 1)
15:    worstFitness ← ∞

16:    for i in 1 : s_tournament do
17:        competitor ← sampleWithoutReplacement(pop, number = 1)
18:        if fit(competitor) > worstFitness then
19:            worstFitness ← fit(competitor)
20:            worstIndividual ← competitor
21:        end if
22:    end for

23:    return worstIndividual
24: end function
```

important thing to note is the relatively large population size, compared with generation GP search heuristics.

Table 4.7: Parameters of the SSOGP search heuristic.

	Variable (Symbol)	Domain	Default
Population Size	mu (μ)	$\mathbb{N}$	300
Tournament Size	tournamentSize ($s_{tournament}$)	$\mathbb{N}$	2
Recombination Probability	recombinationProbabilty (p_{rec})	$[0, 1]$	0.9

4.4.2 Generational Single-Objective Genetic Programming

GSOGP is a single-objective generational GP search heuristic based the very simple $(\mu + \lambda)$ ES algorithm, as described by Like (2013) [195]. In this strategy, μ parents are allowed to breed λ offspring. Then parent and offspring are pooled together, resulting in a pool of $\mu + \lambda$ parents, from which the best μ are selected for the next generation. In state of the art GP systems single-objective search heuristics have been superseded by multi-objective searches, however this search strategy was included mainly as a baseline for comparison and explore the idea that survival analysis could be solved using a more simple single-objective approach.

Algorithm Structure

The classic single-objective generational $(\mu + \lambda)$ evolutionary strategy is implemented in RGP as follows. First, we initialise population $pop(0)$ with μ random individuals, before entering the main evolutionary loop. Next, with random uniform probability and without replacement, we randomly choose $2 \times \lambda$ individuals to act as parents. Then we take the first λ parents to act as mothers, and the second λ parents as fathers. Next we create λ offspring by applying the recombination operator to mothers and fathers. Then we apply mutation operator to the offspring. Next we insert the newly created offspring

into the population to give us a pool of $\mu + \lambda$ individuals to select from. Finally we use rank-based selection to select the best μ individuals form the selection pool, to act as the next generation. This evolutionary process is repeated until the termination criteria is fulfilled. Pseudocode for this simple generational search heuristic is given in Algorithm 4.4. Note that for efficiency reasons RGP would also store the fitness of each individual as it is calculated to avoid recalculating in subsequent iterations.

Algorithm 4.4 Pseudo-code implementation of the GSOGP search heuristic. [83]

```
1:  pop ← createIndividuals (number = μ)

2:  while termination criterion not met do
3:      parents ← sampleWithoutReplacement(pop , number = 2 × λ)
4:      mothers ← parents [1 : λ]
5:      fathers ← parents [ (λ + 1) : 2 × λ]
6:      children ← mut_pop (rec_pop(mothers , fathers))

7:      selectionPool ← parents ∪ children
8:      survivors ← sel_GSOGP(selectionPool , number = μ)
9:      pop ← survivors
10: end while

11: return pop
```

Selection Stratergy

In RGP the selection operator sel_{GSOGP} implements the rank-based selection scheme. Individuals are ranked according to their fitness, and the n highest ranked (best) individuals are selected. This is very simple version of this selection scheme. There are extensions where individuals are assigned a selection probability as a function of their rank in the population. There are a number of such functions proposed, with the linear

and exponential being the most common. For the sake of simplicity these extensions to the rank-based selection scheme were not implemented in this study.

Diversity Preservation

As with previous single-objective search strategy, there is no mechanism to preserve genetic diversity implemented in RGP for this search heuristic.

Parameters

Table 4.8 details the most important parameters appropriate to the RGP implementation of GSOGP. Note that λ should be a multiple of μ and follow the constraint $\lambda \leq (\frac{\mu}{2})$.

Table 4.8: Parameters of the GSOGP search heuristic.

	Variable (Symbol)	Domain	Default
Population Size	mu (μ)	$\mathbb{N}$	100
Children per Generation	mu (λ)	$\mathbb{N}$	50

4.4.3 Generational Multi-Objective Genetic Programming

GMOGP is a generational multi-objective GP search heuristic, based on a classical generational $(\mu + \lambda)$ strategy discussed previously. This approach combines the ideas of multi-objective GP, to control bloat (solution complexity) whilst optimising solutions fitness. A third objective is also added, age, that assists in preserving genetic diversity. This heuristic has been implemented to explore the idea that a multi-objective generational approach can offer improvements over the less complex single-objective approaches in the development of clinical prediction model for censored data.

Algorithm Structure

The classic multi-objective generational $(\mu + \lambda)$ evolutionary strategy is implemented in RGP as follows. First, we apply the initialisation operator to initialise population $pop(0)$ with μ random individuals, before entering the main evolutionary loop. Next, according to the parent selection probability P_{psel}, $2 \times \lambda$ parents are selected either by Pareto selection, or by uniform random sampling without replacement. Then we take the first λ parents to act as mothers, and the second λ parents as fathers. Next we create λ offspring by applying the recombination operator to mothers and fathers. Then we apply mutation operator to the offspring. Next we create ν new individuals using the initialisation operator. Then we insert the newly created offspring and newly created individuals into the population to give us a pool of $\mu + \lambda + \nu$ individuals to select from. Finally we use the Pareto selection operator to select the best μ individuals form the selection pool, to act as the next generation. This evolutionary process is repeated until the termination criteria is fulfilled. Pseudocode for this simple generational search heuristic is given in Algorithm 4.5

Selection Stratergy

In RGP the selection operator sel_{GMOGP} is a Pareto search operator based on NDS with three objectives; fitness, solution complexity, and age (discussed in next paragraph). Crowding distance in the event that there are ties during the NDS. This selection strategy is a kin to the selection strategy of the well-established NSGA-II Evolutionary Multi-Objective Algorithms (EMOA), discussed in section 4.3.4.

Diversity Preservation

GMOGP implements elements of Schmidt & Lipson (2010) Age-Fitness Pareto Optimisation (AFPO) algorithm for preserving genetic diversity and avoiding premature convergence [266]. In each generation, ν newly initialised individuals are inserted into

Algorithm 4.5 Pseudo-code implementation of the GMOGP search heuristic. [83]

pop $\leftarrow$ createIndividuals (number = μ)

while termination criterion not met **do**
  if randomUniformNumber() $\leq p_{psel}$ **then**
    parents $\leftarrow sel_{GMOGP}$ (pop , number = $2 \times \lambda$)
  else
    parents $\leftarrow$ sampleWithoutReplacement (pop , number = $2 \times \lambda$)
  end if
  mothers $\leftarrow$ parents [$1 : \lambda$]
  fathers $\leftarrow$ parents [$(\lambda + 1) : 2 \times \lambda$]
  children $\leftarrow mut_{pop}$ (rec_{pop} (mothers , fathers))
  newIndividuals $\leftarrow$ create Individuals (number = ν)

  selectionPool $\leftarrow$ parents $\cup$ children $\cup$ newIndividuals
  survivors $\leftarrow sel_{GMOGP}$ (selectionPool , number = μ)
  pop $\leftarrow$ survivors
end while

return pop

the population to maintain genetic diversity. These new randomly generated individuals will on average be of low fitness and therefore quickly dominated by older, fitter individuals before having a chance evolve through a series of variation steps. This problem is mitigated by the introduction of genetic age , G, as defined as follows [83]:

$$\begin{aligned} age(g_{new}) &= 0, \\ age[mut(g)] &= age(g) + 1, \\ age[rec(g_A, g_B)] &= max[age(g_A), age(g_B)], \end{aligned} \tag{4.19}$$

where g_{new} is a new genotype just inserted into the selection pool, and g, g_A, and g_B are individuals already existing in a population. New individuals are assigned an age of 0, whilst every mutation operation increments an individuals age by one, and for every recombination operation an individuals age is taken as the largest of its two parent. Solution age is implemented as an objective to minimised, along with solution complexity and fitness

This enables dynamic age-layering of the population, where younger individuals are given a chance to evolve independently of other older, likely fitter and less complex, individuals until they are of the same genetic age (i.e. have undergone a similar amount of variation). It is the hope that his approach will preserve genetic diversity during the GP run, and thus promote exploration of the search, with the ultimate aim of discovering new local optima or even the global optimum.

Parameters

Table 4.9 details the important parameters appropriate to the RGP implementation of GMOGP. Setting the boolean search heuristic parameters *Complexity Control* and *Age Layering* to false will disable the bloat control and diversity preservation objectives, respectively. As in the GSOGP search heuristic and for the same reason, these parameters are subject to the following constraint of $\lambda \leq \left[\frac{\mu}{2}\right]$.

Table 4.9: Parameters of the GMOGP search heuristic.

	Variable (Symbol)	Domain	Default
Population Size	mu (μ)	$\mathbb{N}$	100
Children per Generation	lambda (λ)	$\mathbb{N}$	50
New Individuals per Generation	nu (ν)	$\mathbb{N}_{\not<}$	50
Age Layering	ageLayering	$\mathbb{B}$	true
Parent Selection Probability	parentSelectionP (p_{psel})	$[0, 1]$	1

4.4.4 Termination and Solution Designation

A GP run finishes when a specified *termination criterion* is satisfied. The termination criterion may be a maximum number of generations, maximum run-time, or may be some problem specific success predicate, such as a target fitness. When the run is finished some method of *designating the result* of the run is applied. Typically, the best-so-far individual is identified and designated as the result of the run, although additional individuals and data may be returned as necessary or appropriate to the problem. The RGP package supports the implementation of time, iterations and fitness based termination criterion.

4.5 Genetic Programming Parameters

GP parameters refer to parameters that control the GP run, sometimes referred to as *tuning parameters*, which are typically defined as a preparatory step. There can be a large number of different parameters than can be used to control a GP run dependant on the complexity of the GP system. There are no hard and fast rules about the optimal control parameters that should be used, as these depend too much on the details of the application. A challenge with GP is that, despite having being around for some time, it is a relatively young field and the effects of using various combinations of parameter values is not yet well understood for many applications domains [17]. In their practical advise on GP, Banzhaf et al. (1998) [17] and Poli et al. (2008) [239] both report that in practice GP is robust and likely to work well over a wide range of parameter values and, as a consequence, you do not necessarily need to spend a long time *tuning* GP for it work adequately [239, 17].

An important, arguably the most important, control parameter in GP is *population size*. Other control parameters include *genetic variation rates*, the *maximum solution size*, *maximum number of generations*, and other details of the run. In this section we will discuss these typical control parameters, what is know about their effects , and some

rules of thumb on their use from the literature. However, much of what is known about GP parameters is anecdotal and based on experience of researchers in the field [17].

4.5.1 Population Size

Population size, denoted as μ, is important for a number of reasons. Larger populations take more time and consume more computational resources when evolving a generation. Also, larger populations typically have greater genetic diversity, which increase the search space that can be explored, which in turn may reduce the number of evaluations required for finding a solution. Poli, et al. (2008) [239] suggest that, as a rule one prefers to have the largest population size that your system can handle gracefully; normally, the population size should be at least 500, and people often use much larger populations.

Banzhaf, et al. (1998) [17] report that positive results have been achieved with population sizes ranging from $\mu = 10$ to $\mu = 1,000,000$ individuals, and that in between 10 and $100,000$ individuals they observed a near-linear improvement in performance of the GP system. The authors also state that $\mu = 1,000$ is usually an acceptable starting point for smaller problems and that the population size should grow as the problem grows more difficult. The authors offer a rule of thumb for dealing with more complex problems, that if sufficiently difficult, then the population size should start at and $\mu = 10,000$ individuals and be increased if the other parameters exert heavy selection pressure.

Population size should also be governed by the number of available training cases, with a large number of training cases requiring an increase in μ. Banzhaf, et al. (1998) [17] recommend using $1,000 \leq \mu \leq 10,000$ individuals for between 10 and 200 fitness cases, and $\mu > 10,000$ for more than 200 training cases. Koza reports his experience with population size, using $50 \leq \mu \leq 10,000$ in his book [167], but reports $\mu = 500$ individuals as his most common setting.

4.5.2 Maximum Number of Generation

Some authors suggest limiting the number of generations to 50, arguing that nothing much happens after the fiftieth generation and if a solution hasn't been found by then, its unlikely to found in a reasonable amount of time [239]. However other authors such as Banzhaf et al.(1998) [17] report that they have observed interesting evolution as late as generation 1000, even interesting development occurring as late a generation 10,000.

4.5.3 Primitive Set

Banzhaf et al. [17] offer some rules of thumb for the selection function and terminal sets that they claim have served them well:

- Make the terminal and function sets as small as possible. Larger sets usually mean longer search time
- Its not important to have (all) custom functions in the function set: the system often evolves its own approximations.
- It is very important, however, that the function set contains functions capable of permitting non-linear behaviour, such as if-then functions, boolean operators on numbers, and sigmoid squashing functions
- The function set should be adapted to the problem in the following way: problems that are expected to be solved by smooth curves should use function sets that can generate smooth curves, and problems that are expected to be solved by other types of functions should have at least one representative of these functions in the function set.

4.5.4 Genetic Variarion Rates

The correct balance of crossover and mutations rates is a topic that is wide open in GP, and as such is typically problem specific. Much if the literature suggests that as a starting point, probabilities for applying genetic operations should be very high for crossover ($P_c = 0.9$) and very small for mutation ($P_m = 0.1$), with the mutations rate P_m increased if the results are unsatisfactory. However Poli et al. [239] , and Banzhaf et al. [17] suggest that a different balance ($P_c = 0.5, P_m = 0.5$) may lead to better results for harder problems.

4.5.5 Selection Pressure

Selection pressure is another parameter to consider. In tournament selection, the *tournament size* allow researchers to adjust the selection pressure. A small tournament size causes a low selection pressure and a large tournament size causes high selection pressure. Banzhaf et al. [17] report that they have had very good experiences with low selection pressure, with tournaments of 4 individuals regularly performing well.

4.5.6 Maximum Solution Size

Poli et al. [239] propose that as a rule of thumb, one should try to estimate the size of the minimum possible solution (using the terminals and functions given to GP) and add some percentage (e.g., 50-200%) as a safety margin. Whereas Banzhaf et al. [17] propose that the maximum depth of the trees or the program size should be set such that the programs can contain about ten times the number of nodes as the expected solution size. To allows for error in predicting the solution and for intron growth (discussed in sections 4.6).

Typically the *initial solution size* should be very small compared with the maximum solution size. This enables good solutions to be built up piece by piece using blocks

of in inheritance. However, if this approach is not successful for complex problems, larger initial solutions can be used to add some complexity at the beginning of the GP run to help avoid local optima. A commonly use approach to create a random initial population using the ramped half-and-half approach (discussed in sections 4.3.1), with a depth range of $2 - 6$. These maximum permitted size of initial solutions will be dependent upon the number of the functions, the number of terminals and the arities of the functions. However, evolution will quickly move the population away from its initial distribution [239].

4.6 Bloat: Survival of the Fattest

Most representation schemes in GP, including tree-based GP, allow for variable length solutions. One interesting problem with evolving variable length solutions is the increase of the size of individuals (or solutions) over time. Early on researchers noticed that very often the average size (number of nodes) of the solutions in a population, after a certain number of generations in which it was largely static, at some point would start growing at a rapid pace. Typically the increase in program size was not accompanied by any corresponding increase in fitness [275, 239]

This phenomenon, is commonly referred to as *bloat*. Bloated trees contain a lot of subtrees that don't do anything at all, analogous to superfluous steps in an equation that can be simplified. These subtrees were dubbed *introns*, like their DNA counterparts [194]. Bloat should not be confused with growth. There are times, such as at the beginning of the run where the starting population typically consists of relatively small individuals, where we would expect to see progressive increase in solution size. The distinction between growth and bloat is that for growth we would expect significant increase in size to be associated with a significant increase in fitness. We should therefore define bloat as *solution growth wihtout (significant) return in terms of fitness*.

Bloat is not only surprising, but a real problem for GP. Bloated individuals take longer

to evaluate and use, consume more memory, can be hard to interpret, may exhibit poor generalisation, and are typically far from optimal. Over the years, many theories have been proposed to explain various aspects of bloat, and while great strides have been made, we still lack a single, universally-accepted unifying theory to explain the broad range of empirical observations [239].

Lacking a firm understanding of bloat, GP practitioners have still be faced with the reality of counteracting bloat in their GP runs. As a result a number of ad-hoc, yet effective, approaches have been devised to control bloat. Bloat control has become a very active research area, with several different theories on why bloat occurs and proposed methods for controlling bloat [275, 175, 179, 204, 259, 260]. A review of all these approaches is outside the scope of this thesis. However, we briefly discuss some of the most important.

- Size and Depth Limits
- Anti-bloat Genetic Operators
- Anti-Bloat Selection

The earliest and simplest way to constrain the size of individuals is by placing hard bounds on the maximum allowable size or depth of individuals generated by genetic initialisation and variation operators. This can work in several different ways. Many implementations of this kind of approach generate an offspring and check its size or depth, if its within the pre-specified limits then the offspring enters into the next population, else one of the parents is enters. This type of approach does indeed ensure that solutions grow too large, but has some significant disadvantages. That is, the population is often made up if individuals that almost violate the bounds, which is typically not desired. This problem can be addressed by, rather than return a parent, either classing that evolutionary step as a failure and repeat it, or by returning the oversize offspring but settings fitness to 0. t also well known that depth limits leads to trees that tend to be bushy that are near the depth limit, whereas size limits tend to produce

stringy trees that are close to the size limit. If limits are used they need to be defined in such a way that they do the job of controlling bloat, but at the same time don't constrain the search space so such that suitable quality solutions cannot be found. In section 4.5 we discuss some of the guidance from GP authors on the practical specification of these parameter values.

Bloat can also be controlled by using genetic operators that, directly or indirectly, have an effect on bloat. More recent operators (discussed in sections 4.3.3 and 4.3.2) include *size fair crossover* and *size fair mutation*, which work by constraining choices made during the genetic operation to actively prevent growth. Older methods include mutation operators such as *hoist mutation* and *shrink mutaion*, which help control the average tree size in the population whilst ensuring that new genetic material is introduced.

More recently the trend has been towards penalising large individuals by somehow modulating their selection probability based on their size. This is called the *parsimony pressure method*, which is perhaps the simplest and most frequently used method to control bloat in GP. It works but subtracting a values that based on size from the fitness values of a given individual. Bigger solutions have more subtracted, and thus lowered fitness, which in turn lead to a lower chance of being selected as parents for the next generation. The value to subtract, the penalty, is a function of fitness, size, and a constant known as the *parsimony coefficent*. Some authors have demonstrated some benefits of modifying the value of the parsimony coefficient during the GP run, but most implementation keep this value constant. Recent methods also include the use of *multi-objective optimisation* to control bloat, which typically involves some modified selection mechanism based on the Pareto criterion.

Several variations and extensions of Pareto selection that help control bloat have been proposed in the GP literature. For example, niching via fitness sharing has been proposed to better cover the Pareto front through the inclusion of preference information to focus the selection procedure towards specific regions of the Pareto front [96, 119].

Different applications and variants of Pareto tournament selection [119, 225, 77] as well as the use of other multi-objective optimisation techniques [1, 60, 61] have been developed that control for bloat and/or solutions complexity. Controlling bloat while at the same time maximising fitness turns the evolution of programs into either a multi-objective optimisation problem or, at least, into a constrained optimisation problem [239].

4.7 Implementation

In this section we discuss some of the important consideration when implementing a GP system. We also give an overview of different options including existing GP implementations, as well as options for implementing a GP system from scratch. Finally we introduce the specific GP implementation used for this work and offer justification for implementation choices made.

4.7.1 Implemeting Genetic Programming

When implementing GP there are two main approaches; using existing GP implementations, or implementing a GP system from scratch. There advantages and disadvantages to both options. Implementing from scratch is an excellent way to ensure that you know exactly how the algorithms are implemented and are (typically) easier to customise. However, the downside is the programming expertise required and the need to thoroughly test the system's behaviour. Using an existing implementation is faster and good implementations are often robust, throughly tested, efficient, and well documented. However, heavy customisation can be complicated, with lengthly trial and error, requiring the user to delve into the source code. Also, must but not all existing GP implementation are publicly available, some are commercial and as such come at price. According to Poli et al. (2008), good publicly available GP implementations

include: Lil-GP, ECJ, Open Beagle and GPC++. The most prominent commercial implementation remains Discipulus.

As discussed in chapters 2 and 3, the utility of GP for clinical prediction modelling in the presence of censored data remain unknown. This indicated that significant customisation of the GP implementation was required. For this reason we decided against using an existing implementation of GP, instead opting to build the system from scratch. This was motivated not only by the requirement of heavy customisation by the desire to have full control of the mechanisms involved in the developed GP system.

How GP trees are implemented will obviously depend a great deal on the programming languages and libraries used. Whilst the earliest GP system were implemented in Lisp, people have since coded GP is a huge range of different languages, including C/C++, Java, JavaScript, Perl, Prolog, Mathematica, Pop-11, MATLAB, Fortran, Occam and Haskell [239]. Languages that provide dynamic lists as fundamental data types and appropriate libraries will of course make it easier to implement expression tress and the necessary GP operations.

4.7.2 The R Programming Language

R is a free cross-platform software environment for statistical computing and graphics. It is a GNU project which is similar to the *S language* and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues [247]. R provides a wide variety of statistical and graphical techniques, and is highly extensible. R can be easily extended via libraries, referred to as packages, with more than 5,800 additional packages and 120,000 functions (as of June 2014) available at the Comprehensive R Archive Network (CRAN).

R expressions are internally represented as trees. This makes using R expressions to represent trees a logical choice. R supports the direct manipulation of expression trees through the same syntax for manipulating nested lists, making implementation of

GP operators in R simple, succinct, and easy for proficient R users to understand. R supports *computing on the language*, which greatly simplifies symbolic computation inherent in most GP operations.

As previously discussed, a drawback of building a GP system from scratch is the requirement of proficiency in a suitable programming language. R is no exception with a relatively steep learning curve. A key disadvantage of R is its performance for computationally expensive tasks, of which GP is an example. This arises from two aspects of R, memory and processing. R requires all objects to loaded into memory (RAM),which can be prohibitive on standard-performance computers and large datasets. R is an interpreted language, which provides great flexibility for exploratory data analysis and statistical analysis. However, as an interpreted language R does not offer high performance for computationally expensive tasks, natively working in serial using only a single processing unit. To address this a number of R packages have been developed to address these weakness, such as methods that efficiently manage memory and support parallel computation.

R is freely available and thought its open-source development model, R has an extensive collection packages. As one might expect from a statistical programming language, R has excellent core capabilities for data processing, analysis, statistics, and visualisation. But, though its extensive collection of packages, also provides advanced specialist capabilities for experimental design, machine learning, optimisation, EC, parallel computing, and many others.

4.7.3 RGP: Implementing Genetic Programming in R

Whilst a number of different R packages are typically used in even simple project, there were two packages that offered a significant amount of the functionality required to implement the GP system for this study. These R packages were the publicly available *RGP* and *EMOA* packages.

The RGP package by Flash et al (2014) [84, 83] provides implementations of the initialisation, recombination and mutation operators described in sections 4.3.1, 4.3.2, and 4.3.3, respectively. RGP provides implementation of the single- and multi-objective selection operators described in section 4.3.4. With the multi-objective optimisation being realised through the *emoa* package, which provides functionality such as NDS, Crowding Distance (CD), etc [205]. In addition to search operators, RGP also implements the single- and multi-objective search heuristics described in sections 4.4.1 to 4.4.3. The are also optimised variants of these operators written in C available. See the CRAN package documentation on CRAN for more details [84].

For all experiments in this work it was decided to implement a GP system from scratch using the R statistical programming language and associated packages relevant to GP, namely *RGP* and *EMOA*. The ease of extensibility and customisation were a driving factor in this choice, as a significant amount of development was required to implement GP for survival analysis. The RGP and EMOA, packages provide a substantial amount of functionality required to implement our GP system using R. However, in implementing a GP system for survival analysis there was significant amount of R development required, mainly with respect to the implementation a suitable fitness function for censored data and modification of search operators and heuristics to work with data in the counting process format required by this problem domain, as discussed in section 4.2.2. Another driving factor in using R to implement our GP system was the rich functionality and flexibility that R provides in the ancillary aspects of the GP system, such experimental design, data processing, and the analysis and visualisation of GP results.

4.8 Summary Conclusions

In the first section (4.1) of this chapter we introduced GP, presenting it as an abstract EA for stochastically exploring a vast multidimensional search space.

In section 4.2 we introduced the building blocks of GP , the primitives, and saw how they are combined to produce genotypic representations of individual solutions, that along with GP search operators, define of the search spaces that GP will explore. We also introduced measures of quality, or fitness, of solutions in the search space. We discussed that measures of quality can be multi-facetted, leading to multi-objective fitness. We discussed some of the issues that are presented by censored data and survival analysis, before we defined the fitness measures that will be implemented in this study.

Then in section 4.3 we discussed how the evolutionary process is implemented in GP through the application of search operators, illustrating how these operators define how the GP system will navigate through the set of possible genotypic solutions. We introduced and formally defined initialisation, mutation, recombination, and selection operators, appropriate for tree representations, that will be employed in our subsequent experiments.

In section 4.4 we introduced the set of GP search strategies, grouped by steady-state and generation, that had been selected for this study. We formally defined these search heuristics, including single and multi-objective versions, where appropriate. We also discussed genetic diversity and how its preservation during the GP run is important to its success. We also introduced approaches to control solution complexity and preserve genetic diversity using multi-objective search strategies.

In section 4.5 we look at some of the basic parameters that control a GP run; we summarise the literature to try to understand the effects of these parameters, attempting to offer some 'rules of thumb' on selecting reasonable stating values for these parameters.

In section 4.6 we talk about the phenomenon of intron growth and bloat, how this causes problems in GP, and relate this back to section 4.4 by discussing how certain implementations of genetic operators can counteract this issue.

Finally, in section 4.7 we discussed some of the important consideration when implementing a GP system, giving an overview of different implementation options. Then

we defined the specific GP implementation used for this work and justified the implementation choices made.

In the next chapter, using an observational cohort of patients extracted from CPRD, we will independently and externally validate the performance of the *de facto* cardiovascular risk prediction model for patients with T2DM.

Chapter 5

Experiment 1: External validation of the UKPDS risk engine in incident type 2 diabetes

Next we perform our first set of experiments that aims to address one of the main goals of this work - to motivate the need for improved clinical risk prediction methods. This is achieved by independently and externally validating the performance of the *de facto* cardiovascular risk prediction model for patients with T2DM using a contemporary observational cohort of patients extracted developed from CPRD.

Objective To evaluate the performance of the UKPDS-RE for predicting the 10-year risk of cardiovascular disease endpoints in an independent cohort of UK patients newly diagnosed with type 2 diabetes.

Research Design and Methods This was a retrospective cohort study using routine healthcare data collected between April 1998 and October 2011 from around 350 UK primary-care practices contributing to the CPRD. Participants comprised 79,966 patients aged between 35 and 85 years (388 269 person years) with 4,984 cardiovascular events. Four outcomes were evaluated: first diagnosis of CHD, stroke, fatal CHD, and fatal stroke.

Results Accounting for censoring , the observed versus predicted ten-year event rates were as follows: CHD 6.1% vs 16.5%, fatal CHD 1.9% vs 10.1%, stroke 7.0% vs 10.1%, and fatal stroke 1.7% vs 1.6%, respectively. The UKPDS-RE showed moder-

ate discrimination for all four outcomes, with the concordance-index values ranging from 0.65 to 0.78.

Conclusions The UKPDS stroke equations showed calibration ranging from poor to moderate; however, the CHD equations showed poor calibration and considerably overestimated CHD risk. There is a need for revised risk equations in type 2 diabetes.

5.1 Introduction

National policies for the management of both CVD and type 2 diabetes advocate the calculation of CVD risk in order to identify high-risk patients for targeted interventions [264, 295, 245, 66, 216]. Several multivariable risk-prediction models (or risk scores) have been developed for the general, non-diabetic population that also account for diabetes, but only a few are specific to type 2 diabetes [303]. Only two of these have been developed in patients with newly diagnosed type 2 diabetes, and they both use data from the UKPDS [298]. These two models—one for CHD and the other for stroke—combine to form the UKPDS-RE [165, 284].

International and national clinical guidelines recommended using the UKPDS-RE for predicting cardiovascular risk [216, 134, 32, 218]. Not only is the UKPDS-RE advocated for communicating cardiovascular risk to diabetic patients [295], it has been relied upon for public health decisions [310, 71, 95, 2]. Evidence that these equations are inadequate could bring into question the evidence-base underpinning many clinical decisions and public policies about the management of type 2 diabetes. Two systematic reviews of external validations of type 2 diabetes cardiovascular risk prediction models [303, 35] reported poor calibration of the UKPDS-RE CHD equations in 10 separate studies [103, 139, 283, 322, 155, 237, 58, 304, 280, 302] and differing findings for the stroke equations in two separate studies [155, 58]. The largest of these studies from the UK used only a small sample (n=798) and from a single locality [283]. The largest international study had a larger but still relatively small sample size (n=7

502), using data collated from 20 countries [155].

The purpose of this study was to carry out an external evaluation of the performance of the UKDPS-RE on a large, relatively contemporary dataset of UK-resident patients newly diagnosed with type 2 diabetes.

5.2 Research Design and Methods

This study was carried out using data from CPRD and linked data from the Office for National Statistics and Hospital Episode Statistics. Ethical approval for the study was granted by the CPRD Independent Scientific Advisory Committee on 6th September 2012, protocol number 12_084R (Appendix A).

The CPRD observational data set consists of longitudinal, anonymous records from nearly 700 primary-care practices and more than 11 million patients throughout the UK (based on the January 2012 release) [39]. The computerised data, recorded in the course of routine healthcare by general practitioners and associated staff, included demographic and lifestyle information, medical history, clinical investigations, drug prescriptions, and hospital referrals. Diagnoses in CPRD are recorded using the Read code classification and have been validated in a number of studies, showing a high positive predictive value [118].

Additionally, 357 of the English practices contributing to the data set, representing about 45% of CPRD patients, participate in a linkage scheme by which registered patients are anonymously linked, through a trusted third party, to other, independent data sets [92]. These include hospital-admission data, collated nationally for England as the Hospital Episode Statistics (HES) [68], and mortality data, collated by the Office for National Statistics (ONS) [226]. HES provides details of all National Health Service (NHS) inpatient admissions in England since 1997, including primary and contributory causes coded using the ICD-10 classification. ONS provides details of all deaths in England with immediate and antecedent causes coded using the ICD-9 and ICD-10

classifications. The CPRD dataset and its associated linked datasets are described in more detail in section 2.6.

For this study, a single cohort of patients with incident type 2 diabetes, registered with practices between 1998 and 2011, was identified from the CPRD data set as described below. In order to improve ascertainment of cardiovascular events, only patients whose records linked to the HES and ONS mortality datasets were included, with the former providing details of diagnoses and procedures related to inpatient episodes, and the latter providing both the date and cause(s) of death. The HES data also provided the ethnicity information required for the study. Patients aged between 35 and 84 years at diagnosis were included in the study. As the original UKPDS-RE was based on a cohort aged <65 years, a sensitivity analysis was performed. Patients were excluded if they had on-going or recent CVD (as defined by the UKPDS study criteria), implausible or improbable dates, or missing or indeterminate sex or smoking status. Patients that were HES eligible but had no records in the linked HES data were excluded (n=1727). Patients whose ethnicity was not recorded (n=29,199) were presumed Caucasian and combined with the Caucasian group.

5.2.1 Selection of type 2 diabetes patients

Patients were considered for selection if they had a clinical (Read or ICD-10) code indicative of diabetes mellitus in their CPRD or linked HES records. As not all clinical codes for diabetes distinguish between type 1 diabetes and type 2 diabetes, and some patient histories may erroneously have contained both type 1 and type 2 diabetes codes, these patients were categorized as having type 2 diabetes if they met one or more of the following criteria:

- Clinical codes exclusively indicative of type 2 diabetes
- At least one clinical code indicative of type 2 diabetes (regardless of others indicative of type 1 or non-specific diabetes) and at least one prescription for an

oral hypoglycaemic agent (OHA)

- Prescription of two or more classes of OHA
- Diagnoses of both type 1 and type 2 diabetes and an age of diagnosis older than 35 years.

Any patient with evidence of diabetes secondary to other causes was excluded. The date of diabetes incidence was defined as the date of either first diagnosis or first prescription of a diabetes medication, whichever was earlier. A 'wash-in' period of 365 days was applied to exclude non-incident type 2 diabetes cases.

5.2.2 Outcome measures

The primary outcomes comprised the four cardiovascular events evaluated by the UKPDS-RE: CHD, fatal CHD, stroke, and fatal stroke. To aid comparison, the definition of the outcomes in the CPRD cohort was the same as the definitions from the UKPDS [298, 165, 284]. CHD was defined as the occurrence of fatal or non-fatal myocardial infarction (MI) or sudden death [298]. In patients with multiple CHD events, only the first event was considered. No distinction was made between ischaemic and haemorrhagic strokes. In patients with multiple strokes, only the first stroke was considered. Deaths from causes other than the defined outcomes of interest were treated as censored. Occurrence of clinical events of interest in CPRD were observed from GP-recorded diagnoses, diagnoses recorded during a hospital admission, or cause of death.

5.2.3 Input variables

Values for the input variables required for the UKPDS-RE were taken from CPRD observations around the time of diabetes incidence. Table 5.1 shows the baseline characteristics at the time of incident diabetes. Baseline smoking status was the value re-

corded closest to diabetes incidence, preferring values recorded prior to diabetes incidence; for systolic blood pressure, glycated haemoglobin A_{1c} (HbA_{1c}), total cholesterol, and high density lipoprotein (HDL) cholesterol, the baseline value was the average of biochemical readings recorded in the first two years. The numbers of readings used in deriving these two-year averages were also recorded for use as input parameters (regression dilution) in the UKPDS-RE [284]. Atrial fibrillation was deemed present at baseline if a prior diagnosis or record of a CHADS2 test existed (CHADS2: congestive heart failure, Hypertension, Age $\geq$ 5 years, Diabetes mellitus, prior Stroke or transient ischemic attack).

Table 5.1: Characteristics of patients in the CPRD cohort and UKPDS study. Values are at baseline and are numbers (percentages) unless otherwise stated.

	Females		Males	
Characteristic	CPRD	UKPDS	CPRD	UKPDS
n	36,746	1,879	43,220	2,643
Age (years), mean (SD)	62.6 (12.3)	52.7 (8.7)	60.3 (11.6)	51.5 (8.8)
Ethnicity (%)				
Caucasian/Not recorded	35,452 (96.5)	1,603 (85.0)	42,009 (97.2)	2,151 (81.0)
Afro-Caribbean	404 (1.1)	153 (8.1)	350 (0.8)	201 (7.6)
Asian-Indian	890 (2.4)	141 (7.4)	861 (2.0)	2,91 (11.0)
Smoking status (%)				
Non-smoker	19,684 (54)	—	16,207 (37)	—
Former smoker	10,715 (29)	—	18,173 (42)	—
Current smoker	6,347 (17)	474 (25)	8,840 (20)	898 (34)
Systolic blood pressure (mmHg), mean (SD)*	139 (14)	139 (21)	139 (13)	133 (18)
HbA_{1c} (%), mean (SD)*	7.0 (1.2)	6.9 (1.5)	7.1 (1.2)	6.6 (1.4)
Total cholesterol (mmol/l), mean (SD)*	5.0 (0.9)	5.7 (1.1)	4.7 (0.9)	5.2 (1.0)
HDL cholesterol (mmol/l), mean (SD)*	1.37 (0.32)	1.18 (0.27)	1.17 (0.27)	1.06 (0.23)
Total/HDL cholesterol ratio, mean (SD)*	3.85 (1.03)	—	4.16 (1.11)	—

* Mean of values in the first two years from baseline (HbA1c, systolic blood pressure, and cholesterol)

Multiple imputation was used to replace missing values for systolic blood pressure, HbA_{1c}, total cholesterol, HDL cholesterol, and the number of biochemical readings used in their two-year averages. Multiple imputation is a technique that offers substantial improvements over value replacement approaches based on complete cases or cases matched for age and sex [137]. It involves creating multiple copies of the data and imputing the missing values with plausible values randomly selected from their predicted distribution. Here, we used the Multivariate Imputation by Chained Equations (MICE) library in the R [247] statistical programming language to generate five imputed datasets. Rubin's rules were then used to combine the results from analyses on each of the imputed values, producing estimates and confidence intervals that incorporate the uncertainty of imputed values.

5.2.4 Statistical analysis

For each of the four outcomes, the 10-year estimated risk was calculated for every patient in the CPRD cohort using the UKPDS-RE [165, 284]. Observed 10-year risks were generated using the Kaplan-Meier method, by decile of predicted risk and by five-year age group. The predictive performance of the UKPDS-RE on the cohort was assessed by examining measures of calibration and discrimination.

Calibration refers here to how closely the predicted 10-year cardiovascular risk agreed with the observed 10-year cardiovascular risk. This was assessed for each decile of predicted risk—ensuring 10 equally sized groups—and each five-year age group, by calculating the ratio of predicted to observed cardiovascular risk separately for males and for females. Plotting observed proportions versus predicted probabilities, where a 45° line denoted perfect discrimination, enabled the calibration of the risk-score predictions to be visually assessed.

Discrimination is the ability of the risk score to differentiate between patients who did and did not experience an event during the study period. This measure was quantified

by calculating a concordance index (C-index), in which a value of 0.5 represents random chance, while 1 represents perfect discrimination. All statistical analyses were carried out in R (v2.15.2) [247].

5.3 Results

We identified 79,966 eligible cases, who contributed 383,025, 388,269, 381,833, and 388,004 person years of observed follow-up for CHD, fatal CHD, stroke, and fatal stroke, respectively. The incidence rates for cardiovascular events in the CPRD cohort were 59.2 (95% CI 56.8–61.6), 16.8 (15.5–18.1), 71.2 (68.5–73.2), and 15.2 (14.0–16.5) per 1000 person years for CHD, fatal CHD, stroke, and fatal stroke, respectively. The median durations of follow-up were 4.2 years (inter-quartile range [IQR] 2.0–7.2), 4.3 (2.1–7.3), 4.2 (2.0–7.2), and 4.3 (2.1–7.3), respectively. The proportions of cases followed for 10 years or more were 8.5%, 8.8%, 8.4%, and 8.8%, respectively. Table 5.1 details the characteristics of these patients at or in the first two years from diabetes diagnosis (baseline). People recruited to the UKPDS were a very unusual group of people with type 2 diabetes, and this is reflected in the baseline characteristics. For instance, the mean age at baseline for females in the UKPDS was 53 years versus 63 years in general clinical practice (table 5.1).

5.3.1 Missing data

Complete data on age, ethnicity, smoking status, atrial fibrillation status, systolic blood pressure (SBP), HbA_{1c}, total cholesterol, and HDL cholesterol were available for 70% of females (n=43 741) and 74% of males (n=54 710). Most patients (n=120 572; 88.3%) had missing data on no more than two risk factors (table 5.2). For specific covariates, the proportion of missing data was as follows: HDL cholesterol (26.2% in females, 23.6% in males), SBP (4.1% in females, 3.7% in males), HbA_{1c} (8.1%

in females, 12.0% in males), and total cholesterol (9.5% in females, 12.3% in males) (table 5.3).

Table 5.2: Risk factors used in UKPDS Risk Engine models

Risk Factor	CHD	Stroke
Sex	✓	✓
Age (at diagnosis, in years)	✓	✓
Ethnicity	✓	
Smoking status (at diagnosis)	✓	✓
Atrial fibrillation		✓
Systolic blood pressure (mmHg)	✓	✓
HbA1c (%)	✓	
Total: HDL cholesterol ratio	✓	✓
Duration diabetes (days)	✓	✓

Table 5.3: Completeness of data

No. of risk factors not recorded (per patient)	No. (%) of females (n=36 746)	No. (%) of males 9n= 43 220)
0 (complete data)	43,741 (70)	54,710 (74)
2	10,337 (17)	11,784 (16)
3	2,326 (4)	2,410 (3)
4	723 (1)	644 (1)
5	3,670 (6)	2,970 (4)
6	37 (0)	35 (0)
7	1,648 (3)	1,598 (2)

5.3.2 Discrimination and calibration

A visual illustration of the agreement between mean observed risk and the mean predicted risk, grouped by decile of predicted risk for each of the four UKPDS-RE outcomes is shown in Figure 5.1. Presenting these data in an alternative way, figure 5.2 shows the agreement between the observed risk and the predicted risk by five-year age and sex-specific groups for each of the outcomes. Both the CHD models were clearly miscalibrated—notably for males (overestimating event rates by 174% and 466%, compared with 160% and 398% in females, for CHD and fatal CHD, respectively) and most notably for fatal CHD (overestimating event rates by 440%). There was a clear and consistent over-prediction of risk across all deciles of predicted risk, and across all age and sex-specific groups. The disagreement between observed proportions and predicted risks increased in subsequent deciles of risk and in the older age groups (figures 5.1 and 5.2). The stroke model overestimated event rates by 29% and 58% in females and males, respectively, and the fatal-stroke model underestimated event rates by 20% in males and overestimated these rates by 11% in females. The stroke models showed modest agreement between observed and predicted risk grouped by decile of risk, with the exception of the final, 10th decile for the stroke model in both males and females (figures 5.1 and 5.2). Both the stroke and the fatal-stroke models showed modest agreement across all age groups, with some divergence towards the latter age ranges (70–85 years), most noticeably for males in the stroke model. The fatal-stroke model slightly under-predicted risk for the later age groups, whereas the stroke model tended to over-predict risk for these latter age groups.

Table 5.4 summaries the performance of the four UKPDS-RE models in predicting the 10-year risk in type 2 diabetes patients who were initially free of CVD. The UKPDS-RE overestimated the risk of CHD, fatal CHD, and stroke by 169%, 440%, and 44%, respectively, and underestimated the risk of fatal stroke by 5%. According to the C-index all models were found to have acceptable model discrimination, with the exception of the CHD model in males (C-index=0.65), which was found to have

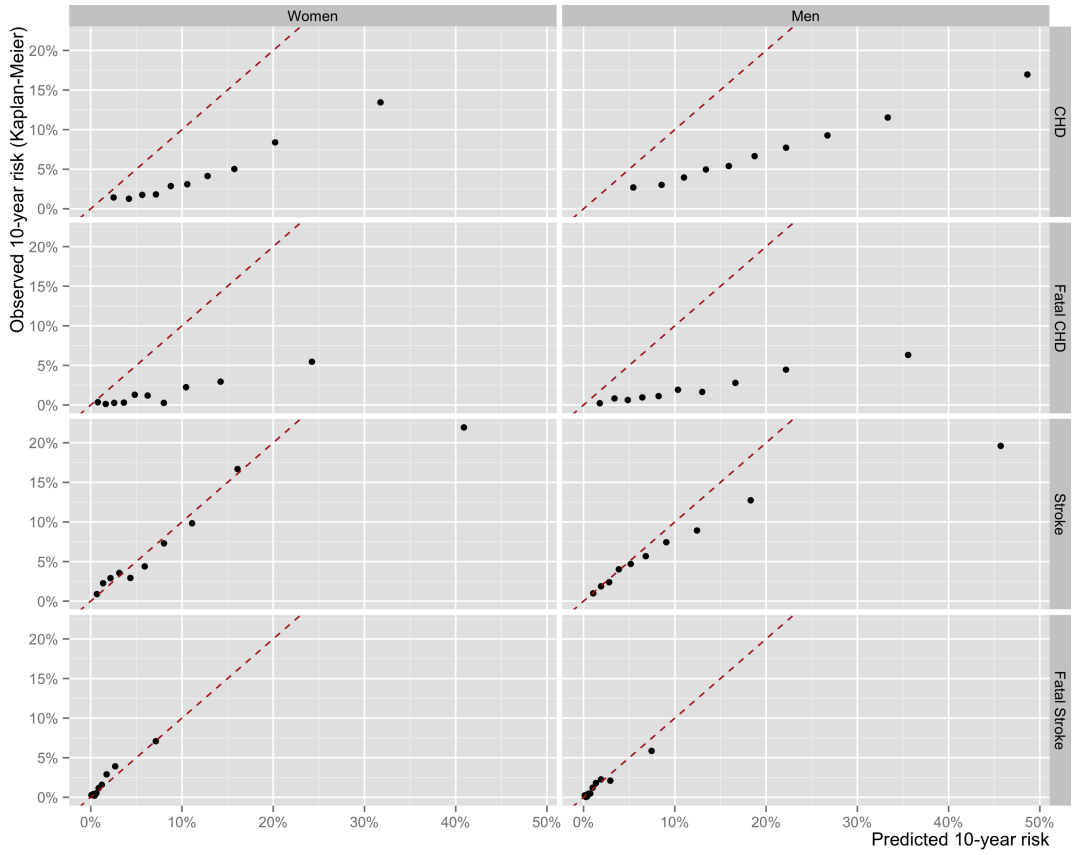


Figure 5.1: Observed versus predicted 10-year risk by sex and outcome

modest discrimination. The C-index values for females and males, respectively, were as follows: for the CHD model, 0.71 and 0.65; for the fatal-CHD models, 0.78 and 0.74; for the stroke models, 0.73 and 0.71; and for the fatal-stroke models, 0.77 and 0.78. All the models showed better discrimination in females, with the exception of fatal stroke, and better discrimination (and variability in estimates) in fatal outcomes in both females and males. Of all the models evaluated, fatal stroke demonstrated the best prognostic separation, with discrimination results ranging from acceptable to good (0.77 and 0.78 in females and males, respectively), whereas CHD exhibited the worst prognostic separation, most noticeably in males, with discrimination results ranging from modest to acceptable (0.71 and 0.65).

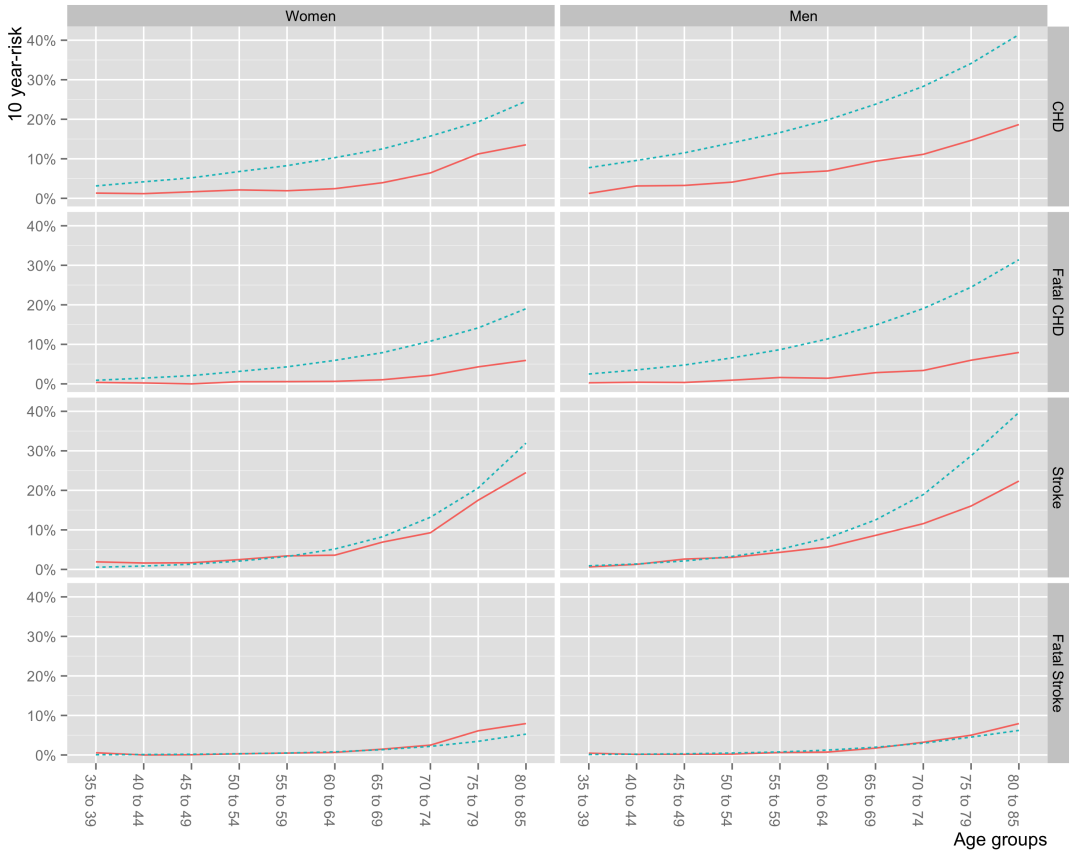


Figure 5.2: Observed and predicted 10-year risks by age group, sex, and outcome (solid lines represent observed proportions and dashed predicted risk).

Table 5.4: Summary of UKPDS-RE performance in predicting 10-year cardiovascular risk

	CHD		Fatal CHD		Stroke		Fatal Stroke	
	Females	Males	Females	Males	Females	Males	Females	Males
n	36,746	43,220	36,746	43,220	36,746	43,220	36,746	43,220
Event rates (%)								
Observed	6.14		1.88		7.00		1.69	
(95% CI)	5.82–6.45		1.70–2.06		6.67–7.33		1.52–1.86	
	4.59	7.44	1.54 2.16	7.28	6.77	1.92	1.50	
	(4.18–5.01)	(6.97–7.90)	(1.29–1.79)	(1.91–2.42)	(6.77–7.79)	(6.34–7.20)	(1.65–2.19)	(1.28–1.72)
Predicted	16.51		10.14		10.10		1.60	
(95% CI)	(16.43–16.59)		(10.07–10.20)		(10.00–10.20)		(1.58–1.62)	
	11.94	20.39	7.66	12.24	9.38	10.71	1.53	1.67
	(11.86–12.02)	(20.31–20.48)	(7.60–7.73)	(12.18–12.30)	(9.28–9.47)	(10.61–10.81)	(1.51–1.54)	(1.65–1.68)
Discrimination (%)								
C-index	0.71	0.65	0.78	0.74	0.73	0.71	0.77	0.78
(95% CI)	(0.69–0.73)	(0.63–0.66)	(0.75–0.81)	(0.72–0.77)	(0.72–0.75)	(0.70–0.72)	(0.74–0.80)	(0.76–0.81)

5.4 Discussion

This validation study showed that the risk equations that constituted the UKPDS-RE were poorly calibrated and significantly overestimated CHD risk. The stroke equations showed calibration ranging from poor to moderate. All the UKPDS-RE equations showed moderate discrimination, with slightly better discrimination for fatal events. This finding was concordant with several other much smaller, external validation studies (<8,000 subjects) that also showed poor calibration and overestimation of CHD risk by the UKPDS-RE [103, 139, 283, 322, 155, 237, 58, 304, 280, 302]. To date, this is the largest study, with around 80,000 patients, and the most comprehensive external validation of cardiovascular-risk prediction in a diverse and more contemporary population with type 2 diabetes.

The relatively poor performance of the UKPDS-RE may be explained, at least in part, by the differences in the baseline profiles of the UKPDS and CPRD populations. These plausibly include: the epidemiological setting, changes in life expectancy, changes in smoking habits, the presence or absence of co-morbidities, temporal changes in diabetes management, and changes in the general quality of care. Other plausible explanations include the possible harm of overly aggressive treatment with sulfonylureas and insulin in the early stages of the disease [54].

The CPRD cohort used in this study was drawn from the UK general-practice, and identified 79,966 patients aged 35–85 years newly diagnosed with type 2 diabetes and registered between 1998 and 2011. The data used to derive the UKPDS-RE risk equations originated from a randomised trial of 5102 UK patients aged 25–65 years newly diagnosed with type 2 diabetes and recruited between 1977 and 1991 (followed up until 1997) [298]. The CPRD cohort comprised patients aged 35–85 for two reasons: patients aged under 35 were excluded to reduce misclassification of type 1 diabetes, and patients aged 66–85 were included to reflect the fact that NICE guidelines recommend using the UKPDS-RE for all ages [295]. Of the 79,966 patients in the CPRD cohort, 31,179 (39%) were outside the 25–65 age range, and a sensitivity analysis suggested

that the inclusion of older subjects aged 65–85 did not significantly affect calibration or discrimination.

A number of the UKPDS trial's exclusion criteria—namely, macrovascular complications, ketonuria, nephropathy, severe retinopathy, malignant hypertension, uncorrected endocrinopathy, and severe concurrent illness—were not applied to the CPRD cohort because their presence would not preclude the use of the UKPDS-RE in clinical practice [298]. It is important to note that, by the nature of trial selection criteria, UKPDS recruits were more likely to be of lower risk, suggesting that the UKPDS-RE would be expected to underestimate risk when applied to the CPRD cohort. Overall, the UKPDS-RE overestimated cardiovascular risk in the CPRD cohort, suggesting that—in spite of the additional exclusion criteria—the UKPDS patients were at higher risk.

A potential difference in the rigour of ascertainment of primary outcomes between UKPDS and CPRD warrants consideration. In this study, we deliberately limited selection to those designated by CPRD as being of research quality, with data linked to HES and ONS mortality data during their entire follow-up period. These criteria combine to make case ascertainment among the highest of any observational data sources. Even prior to the introduction of HES-linked data in CPRD, the predictive value of GP-recorded diagnoses of acute MI in the General Practice Research Database (forerunner to CPRD) exceeded 90% [105].

The secular differences between the UKPDS sample and the current CPRD cohort may have played an important role. The advent of routine diabetes screening in primary care in the UK has almost certainly led to earlier diagnosis of type 2 diabetes than was available at the time of UKPDS recruitment. This is supported by an absolute 2% fall in average incident HbA_{1c} among UK patients with newly diagnosed type 2 diabetes between 1991 and 2012 [127], although the mean HbA_{1c} at specific regimen initiation did not change at all [53]. As such, patients in the UKPDS cohort are likely to have had more advanced diabetes at the point of diagnosis, with correspondingly greater

vascular morbidity.

Over the same period, the diagnosis of MI has evolved from one based solely on clinical symptoms to one that may involve increasingly sophisticated serological and imaging components, such that the severity of MI on admission may plausibly have been reduced. Post-MI care has also improved over the period, and consequently death rates subsequent to MI have fallen. This may partially explain why the UKPDS-RE overestimated fatal CHD, but it does not account for the same discrepancy in non-fatal CHD, which by this rationale could be regarded as conservative.

Another explanation for the disagreement in the observed and predicted risk estimates may be the progressive increase in the use of effective medication for hypertension and dyslipidaemia over the past 20 to 30 years. Of the CPRD patients at baseline (i.e. type 2 diabetes incidence), 22.4% were taking lipid-lowering medication, 49.2% were taking antihypertensive treatment, and 13.7% were taking some form of antiplatelet therapy at baseline. By contrast, the UKPDS was conducted at a time when the number of patients taking such medications was much lower. For example, of the UKPDS patients at baseline, 0.3% used lipid-lowering therapy, 12% used antihypertensive therapy, and 1.6% used more than one aspirin daily. Furthermore, during the period of follow up, less than 2% of UKPDS patients took lipid-lowering therapy at any stage compared with 75.3% of the CPRD cohort [299].

Other changes are also apparent. Only 19% of the CPRD patients were current smokers at baseline, compared with 30% in the UKPDS. The high relative-risk reduction in CHD afforded by statin therapy (subsequent to UKPDS) could have had the effect of reducing the amount of risk that was then potentially modifiable by other interventions such as new glucose-lowering therapies. The benefits of statin therapy are believed to extend beyond their effect on lipid profiles. This is plausible, given that UKPDS-RE considerably overestimated the risk of CHD but not that for stroke. On the other hand, the specific risk markers targeted by these drugs, such as cholesterol, blood pressure, and glucose control are still accounted for within the UKPDS-RE, so the magnitude if

the discrepancy remains difficult to explain. The principal difference between the CHD and stroke models is the presence of HbA_{1c} as an input parameter in the former. The poor calibration of CHD in this study brings into question the role of glucose control in predicting macrovascular complications. In sensitivity analysis—where decile of observed HbA_{1c} was used as the subgroup criterion—there was no gradient in observed risk of CHD, contrary to widespread expectation (figure 5.3). If corroborated, this would have a significant impact on current clinical management guidelines for type 2 diabetes. Our findings might also suggest that, in contemporary practice, the 'benefit' of glucose control (i.e., reduction in CHD risk) is being overstated and consequently is having an undue influence on the diagnosis and treatment of type 2 diabetes.

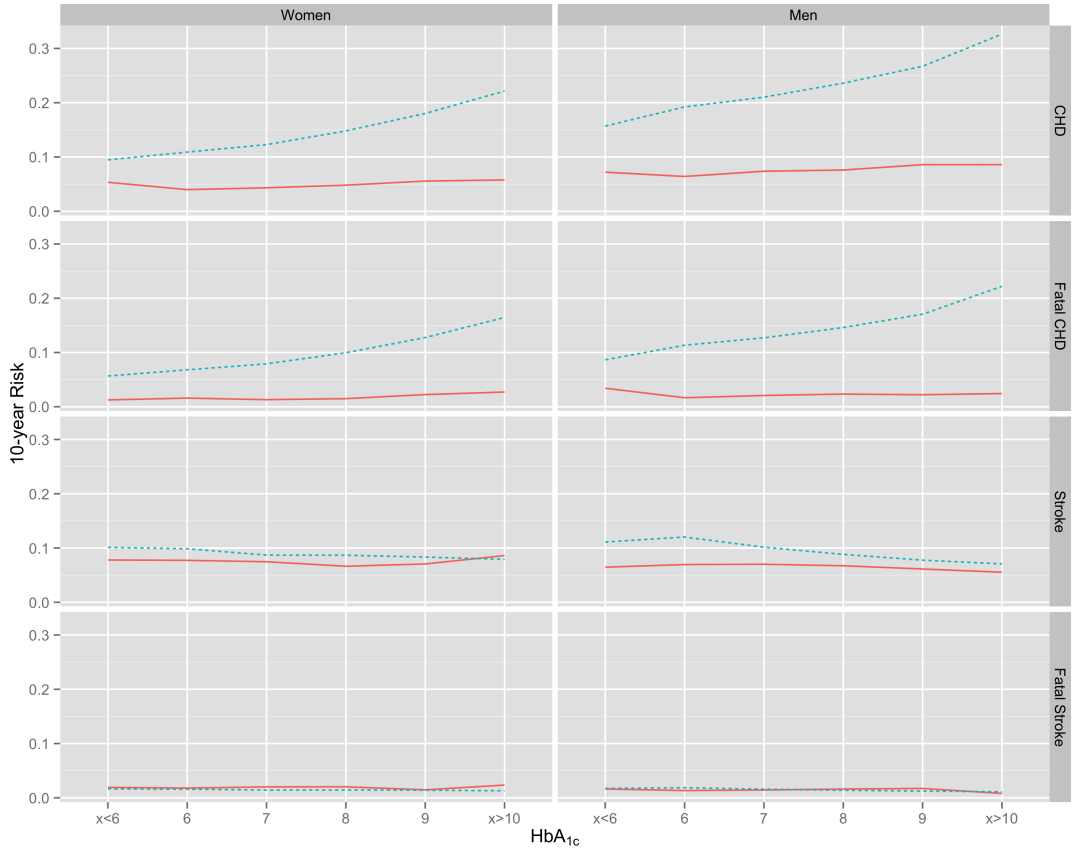


Figure 5.3: Observed and predicted 10-year risks by HbA_{1c}, sex, and outcome (solid lines represent observed proportions and dashed predicted risk).

The overestimation of cardiovascular risk by the UKDPS-RE may also lead to unnecessary targeting of patients for preventative strategies. Accurate estimation of absolute risk is important not only for communicating information on prognosis to patients and practitioners but also for estimating the potential risk-benefit balance, and cost effectiveness of therapy. For example, NICE guidelines for the management of type 2 diabetes recommend using the UKPDS-RE and a specified risk threshold to identify patients not considered to be at high cardiovascular risk for lipid-lowering therapy with statins. Due to the considerable overestimation of cardiovascular risk observed in this study, use of the UKPDS-RE in clinical practice may lead potentially to harmful overtreatment of patients with type 2 diabetes.

5.4.1 Strengths and limitations of the study

A major strength of this study was the size and representativeness of the cohort. Its limitations are the high levels of missing data for HbA_{1c}, total cholesterol, and HDL cholesterol. Omitting cases with missing data and performing a complete-case analysis would have potentially introduced bias into the study. However, the issue of missing data has been addressed by using established methods of multiple imputation. We assumed that people with missing ethnicity data were white. This may have biased the findings to some small degree, but it is unlikely to have impacted substantially on our findings.

Measurement error in identifying the CVD outcomes will have been present in the analysis, but this study has endeavoured to apply the UKPDS study's definitions of the cardiovascular outcomes as far as possible in selecting appropriate medical codes [298]. Moreover, we have supplemented the clinical information recorded in the CPRD with linked but independent secondary-care data from HES, which included details of primary and additional diagnoses for inpatient episodes, and with cause-specific mortality data extracted from death certificates from the ONS. It is therefore unlikely that measurement error is a large source of bias.

Restricting cohort membership to patients from the subset of English practices participating in the linkage scheme between CPRD and HES/ONS should not have introduced significant bias: patient characteristics have been found to be similar between linked and non-linked practices [118]. In order to provide data on ethnicity only those HES eligible patients with a hospital contact were included in our cohort. This excluded only 2% of patients but these patients were presumably healthier than the overall cohort.

Here we have attempted to validate the UKPDS-RE as a prognostic tool in a cohort of newly diagnosed subjects. We did not evaluate its performance with respect to CVD risk among patients with established type 2 diabetes. As the CHD and stroke models each include duration of diabetes as an input parameter, exploration of the utility of UKPDS-RE among prevalent cases of type 2 diabetes is an important future objective.

5.5 Conclusions

The four UKPDS risk equations constituting the UKPDS-RE showed a reasonable ability to identify high-risk patients (discrimination) but were generally poor at quantifying the absolute risk (calibration). The UKPDS-RE CHD risk equations consistently overestimated absolute risk, whereas the UKPDS-RE stroke equations performed relatively well. However, when considered as a whole, the UKPDS-RE was unsuitable for predicting CVD risk in UK subjects with newly diagnosed type 2 diabetes. Our findings suggest that the use of UKPDS-RE in clinical practice will lead to over-estimation of CVD risk in patients with newly diagnosed T2DM. This in turn is likely to lead to selection of preventative treatments, for which, for some patients, the balance of risks may outweigh the benefits. Considering the widespread application of these prediction models in clinical practice, drug reimbursement, and public health decision-making, we suggest that there is a need for revised risk equations in type 2 diabetes. Using the clinical setting of CVD in patients with T2DM to motivate, these findings add support

to the hypothesis that there is a need for improved clinical risk prediction methods.

In the next chapter, we will perform our second set of experiments to demonstrate the utility of genetic programming for the automatic development of clinical prediction models for risk prediction of future cardiovascular events in patients with symptomatic cardiovascular disease using data from the Second Manifestations of ARTerial disease study.

Chapter 6

Experiment 2: A case study in symptomatic cardiovascular disease in the general population using the SMART cohort

Next we perform our second set of experiments to test the main hypothesis—that application of GP can provide more accurate representation of factors that predict the risk of CVD when compared with existing methods—by assessing the utility of the developed GP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events in patients with symptomatic cardiovascular disease using data from the SMART study.

GP is a general methodology, the specific implementation of which requires development of several different specific elements such as problem representation, fitness, selection and genetic variation. Here we implement the specific GP elements developed in chapter 4 to form a GP approach for clinical prediction modelling in the presence of censored survival data, and assess its performance and examine the prognostic significance of different risk factors when compared with the *de facto* statical method in empirical data from the clinical setting of secondary prevention.

Background & Aims Genetic programming is an Evolutionary Computing methodo-

logy, inspired by biological evolution, capable of discovering complex non-linear patterns in large data sets. Despite the potential advantages of genetic programming over standard statistical methods, its applications to survival analysis are at best rare, primarily because of the difficulty in handling censored data. The aim of this study was to demonstrate the utility of genetic programming for the automatic development of clinical prediction models using asymptomatic cardiovascular disease as a case study.

Study Design and Setting We compared genetic programming against the commonly used Cox regression technique in terms of development and performance of a cardiovascular risk score using data from the SMART study, a prospective cohort study designed to identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease. The event predicted was a composite cardiovascular event, comprising of cardiovascular death, non-fatal stroke, and myocardial infarction. The predictive ability of both models was assessed in terms of discrimination and calibration.

Results A total of 3,873 patients were enrolled in the study 1996-2006, aged 19-82 years with a total of 460 cardiovascular events. The study cohort was split 70:30 into derivation and validation sets, used for model fitting and assessment of performance of both the genetic programming and Cox regression models. The discrimination of both models was comparable, albeit in favour in genetic programming; at time points t=1, 3, and 5 years the C-index was 0.65, 0.76, 0.74, and 0.66, 0.70, 0.70, for the genetic programming and Cox regression models, respectively. At the same time points, the calibration of both models was also comparable, but with the Cox modelling better calibrated to the validation data.

Conclusions Using empirical data, we demonstrated that a prediction model developed automatically by genetic programming has predictive ability comparable to that of manually 'tuned' Cox regression. The genetic programming model was more complex but was developed in a fully automated fashion, used fewer predictors as inputs, and did not require the expertise needed for survival analysis. Genetic programming

demonstrated potential as a methodology for the automated development of clinical prediction models for diagnostic and prognostic purposes.

6.1 Introduction

The objective of this study was to develop a tree-based untyped genetic programming approach for censored longitudinal data, comparing it to multi-variable Cox regression in the development of a clinical prediction model for the occurrence of vascular events in patients with symptomatic cardiovascular disease, using data from a prospective cohort study. Four models were developed, three using symbolic regression (GMOGP, GSOGP and, SSOGP) and another using multi-variable Cox regression, and their performance was evaluated in terms of discrimination and calibration in a validation data set.

As discussed in chapter 4, there are a great number of different operators and parameter settings, independent of a particular GP search strategy, that can be used in modern GP systems. The purpose of the experiments in this chapter are to demonstrate the utility of GP for clinical prediction modelling in the presence of censored survival data. As such, we have endeavoured to use 'out of the box', untuned GP, adopting commonly used operators and parameter settings recommended by the literature for the data and type of problem at hand. In this way aim to demonstrate its practical utility by comparing untuned GP, that does not require significant specialist expertise, with highly tuned Cox regression, which does require significant statistical expertise to be applied correctly. Whilst there have been successful applications of genetic programming to regression and classification problems, we believe that this is the first study to have used genetic programming for survival analysis. This serves to demonstrate the utility of genetic programming for the automated development of clinical prediction models.

6.2 Patients and Methods

This study was carried out using data from the SMART study. Details of the ongoing prospective cohort study at the University Medical Centre Utrecht, the Netherlands, designed to identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease have been described previously [276]. Briefly, we consider 3,873 patients who were enrolled in the study between September 1996 and March 2006. Patients were enrolled when presenting at hospital, with follow-up starting from study inclusion. Patients had a clinical manifestation of atherosclerosis, defined as transient ischaemic attack, ischaemic stroke, peripheral atrial disease, Abdominal Aortic Aneurysm (AAA), or coronary heart disease. After informed consent, patients underwent a standardised vascular screening, including a health questionnaire for clinical information, laboratory assessment, and anthropometric measurements at enrolment. During follow-up patients were biannually asked to fill in a questionnaire on hospitalisations and outpatient clinic visits. When a possible event was reported by a participant, correspondence and relevant data were collected (discharge letters, laboratory radiology results). Based on all obtained information, every event was audited by three physicians from different departments.

The primary outcome was any cardiovascular event, comprising of cardiovascular death, non-fatal stroke and non-fatal myocardial infarction (table 6.1). Combing predictor events is a common approach in cardiovascular research to increase statistical power [286]. A cardiovascular event occurred in 460 patients during follow-up.

For our study we *a priori* selected 25 candidate predictors based on previous prognostic studies (Framingham, SCORE). These 25 candidate predictors included risk factors traditionally associated with future events (hyperhomocysteinemia, Intima Media Thickness (IMT) and creatinin level), demographics (age and sex) and risk factors for vascular events in the general population (smoking, alcohol use, BMI, diastolic and systolic blood pressure, lipids and diabetes). Indicators to the location of symptomatic vascular disease (cerebral, coronary, peripheral atrial disease or AAA) and markers of

Table 6.1: Definitions of fatal and non-fatal vascular events in the SMART study

Event	Definition
Ischaemic stroke	Definite: Relevant clinical features that have caused an increase in impairment of at least one grade on the modified Rankin scale, accompanied by a fresh ischaemic infarction on a repeat brain-scan Probable: Clinical features that have caused an increased impairment of at least one grade on the modified Rankin scale; without a fresh ischaemic infarction on a repeat brain-scan
Myocardial infarction	Fatal or non-fatal myocardial infarction: at least two of the following criteria: 1. chest pain for at least 20 min, not disappearing after administration of nitrates 2. ST-elevation $>$ 1 mm in two following leads or a left bundle branch block on the ECG 3. CK elevation of at least two times the value of CK and a MB-fraction $>$ 5% of the total CK
Vascular death	Sudden death: Unexpected cardiac death occurring within 1 h after onset of symptoms, or within 24 h given convincing circumstantial evidence Death from ischaemic stroke Death from intracerebral haemorrhage (haemorrhage on CT-scan) Death from congestive heart failure Death from myocardial infarction Death from rupture of AAA Vascular death from other cause, such as sepsis following stent placement

the extent of atherosclerosis (homocysteine, glutamine, creatinin, albumin, IMT and presence of carotid artery stenosis, table 6.5) were also considered as it is conceivable that they are relevant to predict future events in patients with symptomatic vascular disease. We note that the primary focus of these models is achieving accurate predictions rather than insight into the predictor effects.

6.2.1 Methods

The data set was split, randomly, into two parts: a derivation set of approximately 66.67% (2582 patients) and a validation set of approximately 33.33% (1291 patients). The derivation set was used for model development (both by Cox regression and by genetic programming) and the validation set to access the performance of the two models. The aim for both models was to predict the absolute risk of occurrence of vascular events (stroke, myocardial infarction or cardiovascular death). Given the available follow-up, 1-, 3-, and 5-year risks could be assessed. With respect to sample size in the derivation set, the balance of 313 events and 25 predictors is reasonable, (table 6.5). At least 10-20 events per candidate predictor have been proposed in previous guidelines for the sensible development of predictions models [112, 236, 285, 286].

Multiple imputation is a technique that offers substantial improvements over value replacement approaches based on complete cases or cases matched for age and sex [137]. It involves creating multiple copies of the data and imputing plausible values randomly selected from their predicted distribution. Here, we used multiple imputation to replace missing values for smoking status, packyears, alcohol, BMI, diabetes, SBP, Dyastolic Blood Pressure (DBP), Total Cholesterol (TC), High-density Lipoprotein Cholesterol (HDL), Low-density Lipoprotein Cholesterol (LDL), triglycerides, homocysteine, glutamine, creatinine, albumin, IMT and carotid artery stenosis (table 6.5), generating five imputed data sets. The first set of imputations were used for further analysis ('single imputation'). Although multiple imputation is preferable from a theoretical view point, single imputation was considered more practical and sufficient to

obtain reasonable predictions [286]. Final models were also constructed with multiply imputed data sets to check for any relevant differences in point estimates, and widening of confidence intervals.

Cox Regression

The SMART data set, like many others in cardiology and oncology, is an example of censored data, often referred to as a survival outcome. In medical and epidemiological studies the Cox Proportional Hazards model (or Cox regression) is the most often used model for survival outcomes [48]. Analogous to this model for a binary outcome in uncensored data, where we know whether or not the patient experienced the event in the time horizon of interest, is the logistic model. Multi-variable logistic regression model is is the most widely used statistical technique nowadays for binary medical outcomes [107, 286, 307]. The Cox PH model is the natural extension of the logistic model to the survival setting [286].

In the derivation set, we fitted a Cox regression model using a similar modelling strategy that described by Steyerberg [286] in the development of a clinical prediction model on the SMART study data set. Briefly, the we first fitted a full main effects model. Biologically implausible values were set to missing (prior to imputation) and extreme values truncated at the 1st and 99th centile. To enhance the flexibility of the Cox regression and enable fairer comparison with the (unrestricted) genetic programming, we considered continuous predictors (e.g. age, creatinine, blood pressure) for transformation. Several transformations were considered in adding polynomials, fractional polynomial terms, transformations (e.g. log, square root, exponential), restricted cubic splines (with varying number of knots) and linear coding (i.e categorisation). To further enhance a fair comparison with genetic programming, we considered interaction effects between predictors. Key limitations of the Cox PH model include the assumption of proportional hazards - that hazard functions in the different strata are proportional over time, assumptions of linearity and additivity which are implicit in re-

gression's linear combinations, and the fact that the baseline hazard is never specified (although this last one may be advantage in some cases). All model assumptions relevant to the Cox proportion hazards model were tested. A reduced model was obtained by applying a backwards selection procedure, with Akaike information criterion (AIC) as the stopping criterion.

Internal validation of the model was performed using a bootstrapping re-sampling procedure [22, 75, 111]. Random samples were drawn (with replacement) from the derivation set with 200 replications, and the backwards selection of predictors for the reduced model repeated each time. Bootstrapping yielded an estimate of optimism of the reduced models as expressed by the concordance (C) statistic, which for a binary outcome is identical to the area under the receiver operating characteristic (ROC) curve. A shrinkage factor was derived from the bootstrap estimates to re-calibrate the model to adjust for optimism. The re-calibrated model was applied to the validation set to estimate its discrimination and calibration in an independent sample. All analyses were carried out in R (v3.0.1) [247].

Symbolic Regression

For the experiments in this chapter we implemented three untyped tree-based GP models using steady-state single-objective (SSOGP), generational singe-objective (GSOGP) and, generational multi-objective (GMOGP) search strategies, discussed in sections 4.4.1 - 4.4.3, to fit symbolic regression models to the data to estimate discrete hazard, thus predicting the risk for cardiovascular events. Here the outcome is discrete hazard rate which is the conditional probability that an individual will experience the event during time interval [t-Δ/2, t+Δ/2), given they are event-free at the beginning of the interval, as opposed to continuous hazard rate in the Cox regression. We have modelled survival in discrete time, as opposed to continuous time, to take advantage of the discrete-time survival fitness function (Equation 4.13) detailed in section 4.2.3, which enables GP to be applied to censored survival data. An advantage of this model is that it's not

constrained by the assumption of proportional hazards and is better suited to any non-linear interactions between explanatory variables. However, the data need to prepossessed into the 'counting process format', which results in multiple observations per subject, representing the discrete time segments for which they were observed.

Symbolic regression was performed using the RGP package in the R statistical programming language, with the SSOGP, GSOGP and, GMOGP search heuristics using the parameter settings detailed in tables 6.2 - 6.4. The choice of these starting parameters was driven by the size of the training data available, the perceived relative complexity of the problem and, the recommendation and guidelines from authors in the field of GP which are discussed in section 4.5. Wherever possible we have opted for the most common or default operators and parameter settings, opting not to tune the parameters.

The default RGP function set ($+, -, \div, \times, \sin, \cos, \tan, \sqrt{}, \exp, \log$) was used to enable the representation of potential non-linear relationships present in the training data. Koza's [167] ramped half-and-half random initialisation method, the most commonly used initialisation operator in tree-based GP, was used with a maximum tree depth of 63. The GP approach utilised is untyped with the search space constrained by the use of fitness penalties. Specifically, the 'death penalty' is used where invalid solutions, such invalid mathematical operations (e.g. dividing by 0), have their fitness value set to ∞ giving them the lowest possible fitness (section 4.2.1). As discussed in section 4.5, some authors propose that as a rule of thumb to specifying the maximum tree depth, one should try to estimate the size of the expected solution size and add some percentage as a safety margin. For this experiments we calculated, when transforming categorical predictors into 'dummy' variables (i.e. n-1 dummies per categorical predictors, where n is the number of levels), that the expected solution depth would be 21 based on the expected solution being modelled as a regression model with 19 predictors. Based on this we estimated a $maxdepthsize = 21 \times 3\% = 63$. In most real-world GP applications only a fixed compute time budget is available. Therefore, the expiration of a fixed time compute budget was chosen as the termination criterion.

The compute time budget for these experiments was set to 12 hours wall-time and these experiments were run on a single thread on an Intel Westmere 2.8GHz CPU with 48 GB of memory. Because larger population sizes tend to increase genetic diversity, because this problem is a relatively difficult one and, the data has a relatively large number of training cases, we wanted to use the largest population size that the GP system could handle gracefully. Based on observations from practitioners (see section 4.5) in the filed we opted for an initial population size of $\mu = 1,000$. The most commonly used mutation and recombination operators in tree-based GP, Subtree Mutation (Equation 4.15) and Subtree Crossover (Equation 4.18), were selected for this experiment with the default RGP parameter settings. Because this is relatively hard problem we have opted genetic variation rates of 0.5 and 0.5 for crossover (P_{rec}) and mutation ($1 - P_{rec}$), respectively. For the SSOGP search heuristic implemented in these experiments we have opted for a low selection pressure ($tournamentsize = 4$) because, as discussed in section 4.5, authors have have very good experiences with low selection pressure, with tournaments of 4 individuals regularly performing well.

Table 6.2: Parameters of the SSOGP search heuristic.

	Variable (Symbol)	Domain	Setting
Population Size	mu (μ)	$\mathbb{N}$	1,000
Tournament Size	tournamentSize ($s_{tournament}$)	$\mathbb{N}$	4
Recombination Probability	recombinationProbabilty (p_{rec})	$[0, 1]$	0.5

Table 6.3: Parameters of the GSOGP search heuristic.

	Variable (Symbol)	Domain	Setting
Population Size	mu (μ)	$\mathbb{N}$	1,000
Children per Generation	mu (λ)	$\mathbb{N}$	500

We did not perform internal validation in the genetic programming approach using a bootstrap as we did with the Cox regression, because it would not have been possible to convert it into a shrinkage factor in the same way as we would for a Cox regression.

Table 6.4: Parameters of the GMOGP search heuristic.

	Variable (Symbol)	Domain	Setting
Population Size	mu (μ)	$\mathbb{N}$	1,000
Children per Generation	lambda (λ)	$\mathbb{N}$	500
New Individuals per Generation	nu (ν)	$\mathbb{N}_{\not<}$	500
Age Layering	ageLayering	$\mathbb{B}$	true
Parent Selection Probability	parentSelectionP (p_{psel})	$[0, 1]$	1

The genetic programming system is a stochastic process, with each run potentially yielding models with differing complex structures (i.e. symbolic regression). As a result regression coefficients do not exist in genetic programming models in the same way that they do in regression models. Instead the training data was split 2/3:1/3 into training and holdout sets, using a stratified random split to ensure proportionate number of events. The first 2/3, the training set, was used for training to induce a population of prediction models. The remaining 1/3, the holdout set, was used at the end of the genetic programming run to calculate the fitness of the population of models and thus determine the fittest or 'best of run' model to be returned as the output of the genetic programming system. In this way the final genetic programming model was selected based on its fit to unseen data using a sample other than which it was trained or developed.

To understand variable selection in the genetic programming and enable comparison with bootstrapped backwards selection of the Cox model, the genetic programming system was executed 25 times to produce 25 'suggested' models. For each iteration the training data was randomly (stratified) split 2:1. The final genetic programming model was applied to the validation data set to assess its performance, in terms of discrimination and calibration in an independent sample. All analyses were carried out in R version 3.1.2 [247].

Comparison of both methods

The four clinical prediction models, one obtained from Cox progression and three from symbolic regression (SSOGP, GSOGP and, GMOGP), were evaluated in terms of overall survival curves, discrimination and calibration in the validation data set. The models were used to predict the discrete hazards $h(t)$ at t = 1, 3, and 5 years. Model were first evaluated visually by comparing the survival probabilities $S(t)$ predicted by the models with estimates obtained using the KM method. The agreement between these curves and the KM estimates were assessed visually.

Discrimination is the ability of the risk score to differentiate between patients who did and did not experience an event during the study period. This measure was quantified by calculating a concordance statistic (C-statistic), proposed by Harrell et al. [109, 110, 111, 300] which is a rank-based measure for censored survival data. The C-statistic is equivalent of the AUC measure [107] for survival data, in which a value of 0.5 represents random chance and 1 represents perfect discrimination. The C-Statistic was evaluated considering truncation of the survival/censoring times at t=1, 3, and 5 years.

Calibration refers here to how closely the predicted x-year cardiovascular risk agreed with the observed x-year cardiovascular risk. Model calibration was assessed using calibration plots and the generalisation of the Hosmer-Lemeshow test statistic for survival data [56]. This was assessed by grouping subjects into g equally sized groups, with the same cardinality, based on quantiles of predicted $S(t)$, where t is a fixed time point, and calculating the ratio of predicted to observed cardiovascular risk. For each of the g groups, plotting observed proportions (KM) versus predicted probabilities (model) enabled the calibration of the model predictions to be visually assessed. The closer the g points are to the 45 degree line connecting (0,0) to (1,1), the better is the calibration. To obtain the χ^2 statistic, the model predicted number of events was calculated, for each group, as the product of the group size by the average predicted incidence $1 - S(t)$. The results were then compared to the observed number of events in the corresponding

groups calculated as the product of the group size by the KM estimate of $1 - S(t)$. This leads to a statistic which, under the null hypothesis of numerical agreement between predicted and observed number of deaths, has a χ^2 distribution. Calibration was evaluated by grouping subjects according to the predicted $S(t)$ at t = 1, 3, and 5 years. All analyses were carried out in R version 3.1.2.

6.3 Results

6.3.1 Descriptives

There were no major differences in the baseline characteristics of the patients between the derivation and validation sets (table 6.5). Data were available on 9,636 and 4,895 person-years collected during a median follow-up of 3.3 (range, 0-9 years) and 3.3 years (0-9 years) for the derivation and validation sets, respectively. In the derivation set a total of 313 events occurred, corresponding to 1-, 3-, and 5-year cumulative incidences of 4.1%, 8.9% and 15.0% respectively. In the validation set a total of 147 events occurred, corresponding to 1-, 3-, and 5-year cumulative incidences of 3.8%, 8.1% and 12.0% respectively.

6.3.2 Model Derivation

Prior to modelling extreme values in IMT, BMI, lipids (cholesterol, HDL, LDL, triglycerides), homocysteine and creatinine were truncated at the 1st and 99th centile. Indicators to the location of symptomatic vascular disease (cerebral, coronary, peripheral atrial disease or AAA) were optimally combined into a single variable (or sumscore), with each condition contributing one point except AAA that contributed 2 points. Using univariate cox models there was no significant difference between the sumscore (χ^2 119; 1 d.f.) and using the separate terms (χ^2 123; 4 d.f.) however there was a saving of 3 degrees of freedom from the sumscore.

Table 6.5: Baseline characteristics of patients in the SMART cohort, by derivation and validation sets (n=3,873).

		N	Test set $N = 1291$	Training set $N = 2582$	Test Statistic
Cardiovascular event		3873	11% (147)	12% (313)	$\chi^2_1 = 0.45,\ P = 0.51^1$
Gender : Female		3873	25% (320)	25% (656)	$\chi^2_1 = 0.18,\ P = 0.68^1$
Age	years	3873	52 60 68	52 60 68	$F_{1,3871} = 0.03,\ P = 0.86^2$
Smoking : Never		3873	18% (235)	18% (458)	$\chi^2_3 = 5.6,\ P = 0.13^1$
Former			69% (885)	71% (1826)	
Current			12% (158)	11% (286)	
NA			1% (13)	0% (12)	
Packyears	years	3852	5.2 18.2 33.8	6.1 19.5 34.5	$F_{1,3850} = 0.79,\ P = 0.38^2$
Alcohol : Never		3873	20% (255)	19% (496)	$\chi^2_3 = 1.1,\ P = 0.77^1$
Former			11% (141)	10% (267)	
Current			69% (885)	70% (1804)	
NA			1% (10)	1% (15)	
Body mass index	Kg/m2	3870	24 26 29	24 26 29	$F_{1,3868} = 3,\ P = 0.084^2$
Diabetes : 0		3873	76% (983)	78% (2004)	$\chi^2_2 = 1.1,\ P = 0.59^1$
1			23% (294)	21% (552)	
NA			1% (14)	1% (26)	
Systolic blood pressure, automatic	mm Hg	2650	127 140 155	127 139 153	$F_{1,2648} = 1.4,\ P = 0.23^2$
Diastolic blood pressure, automatic	mm Hg	2652	73 79 86	73 79 86	$F_{1,2650} = 0.01,\ P = 0.9^2$
Systolic blood pressure, by hand	mm Hg	2375	128 140 158	125 139 155	$F_{1,2373} = 3.8,\ P = 0.052^2$
Diastolic blood pressure, by hand	mm Hg	2374	75 82 90	74 82 90	$F_{1,2372} = 0.2,\ P = 0.65^2$
Total cholesterol	mmol/L	3855	4.4 5.2 5.9	4.3 5.1 5.9	$F_{1,3853} = 2.6,\ P = 0.11^2$
High-density lipoprotein cholesterol	mmol/L	3843	0.95 1.15 1.40	0.97 1.18 1.43	$F_{1,3841} = 3.8,\ P = 0.05^2$
Low-density lipoprotein cholesterol	mmol/L	3657	2.5 3.1 3.8	2.4 3.0 3.8	$F_{1,3655} = 3.2,\ P = 0.073^2$
Triglycerides	mmol/L	3845	1.1 1.6 2.3	1.1 1.5 2.2	$F_{1,3843} = 4.1,\ P = 0.042^2$
Cerebral		3873	30% (387)	29% (760)	$\chi^2_1 = 0.12,\ P = 0.73^1$
Coronary		3873	56% (724)	56% (1436)	$\chi^2_1 = 0.08,\ P = 0.78^1$
Peripheral		3873	24% (308)	24% (632)	$\chi^2_1 = 0.18,\ P = 0.67^1$
Adominal aortic aneurysm		3873	10% (134)	11% (282)	$\chi^2_1 = 0.26,\ P = 0.61^1$
Homocysteine	(μ)mol/L	3410	10 13 16	10 13 16	$F_{1,3408} = 2.5,\ P = 0.11^2$
Glutamine	(μ)mol/L	3854	5.3 5.8 6.5	5.3 5.7 6.5	$F_{1,3852} = 0.94,\ P = 0.33^2$
Creatinin	mL/min	3856	78 89 102	78 89 101	$F_{1,3854} = 0.62,\ P = 0.43^2$
Albumin : No		3873	75% (969)	75% (1928)	$\chi^2_3 = 1.1,\ P = 0.78^1$
Micro			17% (221)	17% (434)	
Macro			3% (33)	3% (81)	
NA			5% (68)	5% (139)	
Intima media thickness	mm	3775	0.75 0.88 1.05	0.75 0.88 1.07	$F_{1,3773} = 0.24,\ P = 0.63^2$
Presence of carotid artery stenosis : 0		3873	79% (1020)	79% (2038)	$\chi^2_2 = 0.91,\ P = 0.63^1$
1			18% (236)	19% (486)	
NA			3% (35)	2% (58)	

$a\ b\ c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. N is the number of non–missing values. Numbers after percents are frequencies. NA represent missing values. Tests used:[1]Pearson test; [2]Wilcoxon test

The full Cox regression model consisted of 14 predictors, several of which had limited contributions. Predictors that had a relatively large effect were age, location of symptomatic vascular disease (sum score), albumin, and the marker of renal damage, creatinine. The coding of predictors that gave the best representation for age and creatinine were $(AGE - 50)^2$ and log(CREAT), respectively. We also tested interactions between predictors but the resultant interactions were not considered relevant enough to include any interaction terms in the final model. The proportionality of hazards was tested using an overall test which was not significant. We judged our sample size to large enough to allow for some model reduction (313 events and a full model with 17 degrees of freedom), facilitating easier practical application and clinical interpretation. We applied a backwards step-wise selection procedure, using AIC as the stopping rule, to achieve a reduced Cox model. The reduced step-wise selected model was found to be optimal with 9 predictors (table 6.6). Predictors with relatively weaker effects (alcohol, diabetes, gender, smoking status, and stenosis) were excluded from the reduced model.

Bootstrapping of the reduced model yielded an estimate of required shrinkage for the coefficients in the step-wise selected model of 0.91, suggesting that each coefficient should be reduced by 9% to obtain a re-calibrated model that corrects for optimism. This shrinkage factor was applied to the reduced backwards step-wise model and considered the calibrated 'final' Cox regression model (table 6.7). All analyses were repeated with the multiply imputed data sets, with largely similar results.

Based on the parameter settings detailed in section 6.2.1, the generational approaches failed to complete within the allocated 12 hour fixed time budget because they exceeded the maximum memory allocation (42Gb). This is likely to be due to the increased computational expense of the generational approach to GP, when compared to the computationally cheaper steady-state approach. The experiments were repeated with a range of different parameter settings that effect the memory requirements, such as population size and max depth of solutions, but unfortunately these all exceeded the

Table 6.6: Cox regression coefficients in the full model, and stepwise selected model (using AIC).

Predictor		Full	Stepwise
Age	AGE	0.0011	0.0011
Albumin	ALBUMIN=Macro	0.5289	0.5371
	ALBUMIN=Micro	0.5227	0.5184
Alcohol	ALCOHOL=Current	0.0234	
	ALCOHOL=Former	−0.1854	
Body mass index	BMI	−0.0383	−0.0359
Creatinin	CREAT	0.5992	0.5282
Diabetes	DIABETES	0.0783	
High-density lipoprotein cholesterol	HDL	−0.4619	−0.4096
Previous atherosclerosis (sum score)	HISTCAR2	0.2980	0.2895
Homocysteine	HOMOC	0.0169	0.0182
Intima media thickness	IMT	0.5145	0.5879
Gender	SEX=Female	0.1754	
Smoking	SMOKING=Current	0.0798	
	SMOKING=Former	0.0427	
Presence of carotid artery stenosis	STENOSIS	0.1815	
Systolic, by hand	SYSTH	0.0037	0.0041

available memory and failed. From here on we will discuss only the results from the SSOGP symbolic regression models.

Figure 6.1 describes the run statistics for the 25 GP SSOGP runs performed on the different stratified re-samples of the derivation data set. The figure depicts the evolution of the different GP run's best fitness (bestFit) and complexity, quantified using

Table 6.7: Association of each predictor with cardiovascular events in the calibrated final Cox model.

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
AGE	52.00	68.0	16.00	0.32	0.08	0.16	0.48
Hazard Ratio	52.00	68.0	16.00	1.38		1.18	1.61
BMI	24.03	28.7	4.69	-0.15	0.08	-0.31	0.00
Hazard Ratio	24.03	28.7	4.69	0.86		0.73	1.00
SYSTH	127.00	156.0	29.00	0.11	0.07	-0.04	0.25
Hazard Ratio	127.00	156.0	29.00	1.11		0.96	1.29
HDL	0.96	1.4	0.47	-0.18	0.09	-0.34	-0.01
Hazard Ratio	0.96	1.4	0.47	0.84		0.71	0.99
HISTCAR2	1.00	5.0	4.00	1.05	0.27	0.52	1.59
Hazard Ratio	1.00	5.0	4.00	2.87		1.67	4.91
HOMOC	10.50	15.9	5.40	0.09	0.05	-0.02	0.19
Hazard Ratio	10.50	15.9	5.40	1.09		0.98	1.21
CREAT	78.00	101.0	23.00	0.12	0.05	0.04	0.21
Hazard Ratio	78.00	101.0	23.00	1.13		1.04	1.24
IMT	0.75	1.1	0.32	0.17	0.07	0.04	0.30
Hazard Ratio	0.75	1.1	0.32	1.19		1.04	1.35
ALBUMIN — Micro:No	1.00	2.0		0.47	0.14	0.21	0.74
Hazard Ratio	1.00	2.0		1.60		1.23	2.09
ALBUMIN — Macro:No	1.00	3.0		0.49	0.24	0.02	0.96
Hazard Ratio	1.00	3.0		1.63		1.02	2.61

mean visitation length (meanVisLen), over time. Time is represented as iterative or evolutionary steps (stepNo), where an evolutionary step is each time that tournament selection is performed and new individuals are generated and considered for inclusion in the population. The 'final' symbolic regression model is the individual the best (i.e. lowest) fitness at the end of all 25 GP runs. We can see that there is significant vari-

ation in the best fitness and complexity of the individuals developed by the different GP runs. However, in general the improvement in best fitness appears to level out towards the over time in all runs, suggesting that the selected fixed time compute budget of 12 hours wall time is acceptable. A range of different run statistics for all 25 SSOGP runs are detailed in Appendix B.

The final model produced by genetic programming model included 6 predictors: age (AGEn), sum score of previous atherosclerosis (HISTCAR2n), gender (SEXfemale.n), IMT (IMTn), homocysteine (HOMOCn), and albumin (ALBUMINNo.n), in addition to the discrete time indicator (t_j), which is present in all the genetic programming models to represent the jth time interval. The final prediction model generated by genetic programming is presented in figure 6.2, which is a binary parse tree representing Equation 6.1.

$$
\begin{aligned}
\hat{\lambda}(t_j, X) =& Prob(T = t_j | T \geq t_j, X) = \frac{1}{1 + e^{-X\hat{\beta}}}, where \\
X\hat{\beta} =& (t_j - (0.441 + t_j)) * \exp(\sin(\sin(ALBUMINNo.n))) * \\
& \exp((HOMOCn + AGEn)/\tan(1.889)) * \cos((\tan(SEXfemale.n) \\
& + (HOMOCn + AGEn))/\exp(\cos((\tan(\tan(\exp(HISTCAR2n))) \\
& + \sin(IMTn) + \sin(IMTn))/\tan(1.886)))) * 2.487 - \exp(\cos(\\
& HISTCAR2n/\tan(\tan(-1.813))/\tan(\tan(\tan(0.739)))))
\end{aligned}
\tag{6.1}
$$

The other 24 prediction model generated by genetic programming are presented in Appendix C.

The genetic programming approach was applied 25 times, each time trained and tested on a different stratified re-sample of the derivation data set. This leads to a pool 25 different 'best of run' models, each of which may have selected different subset of predictors as inputs and as such may have differing levels of performance. In this pool of

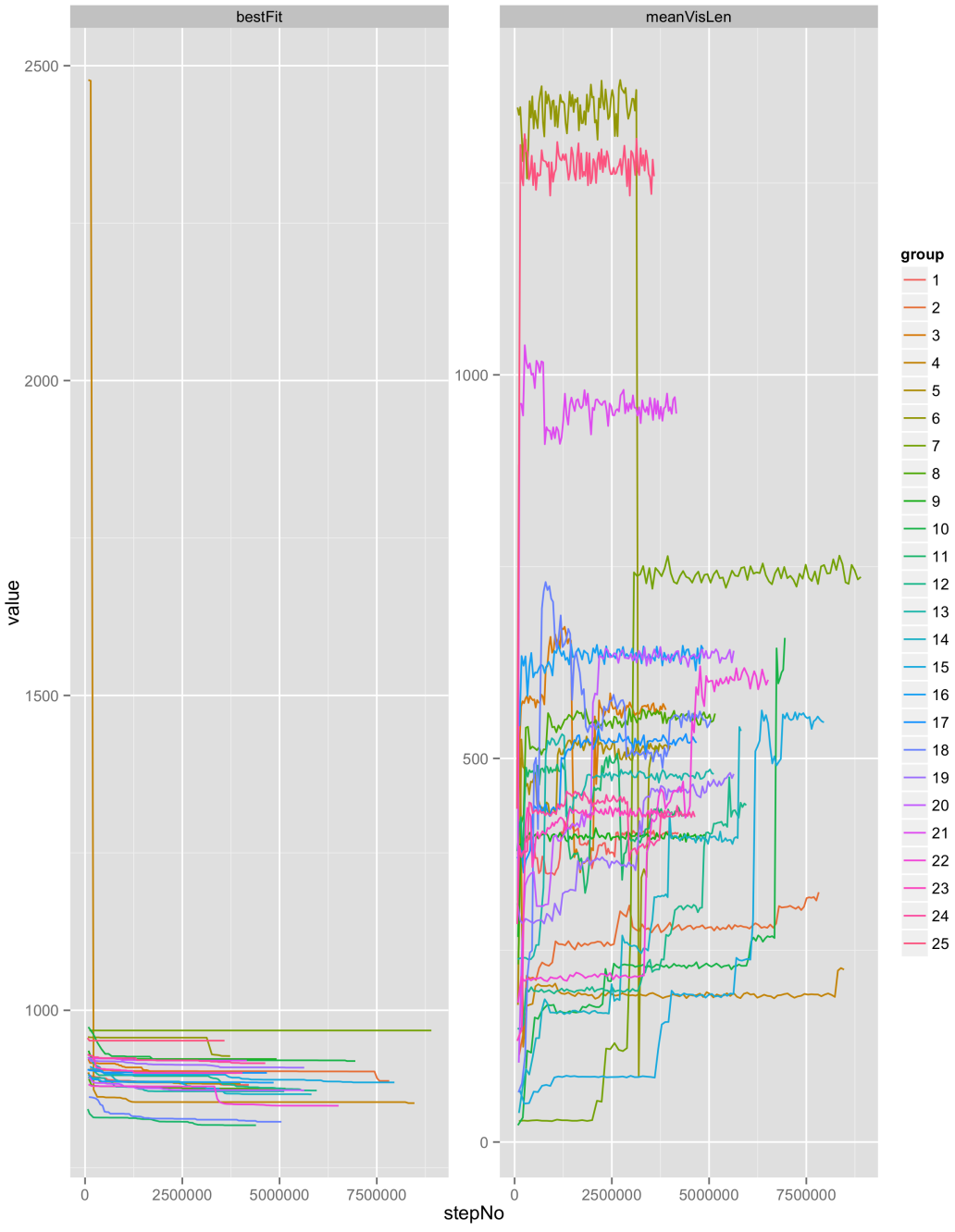


Figure 6.1: Selected runs statisitcs for the 25 SSOGP runs in the SMART experiments.

genetic programming models, the mean number of predictors used was 6 (IQR: 5—8). The backwards step-wise selection procedure used in the Cox modelling was also re-

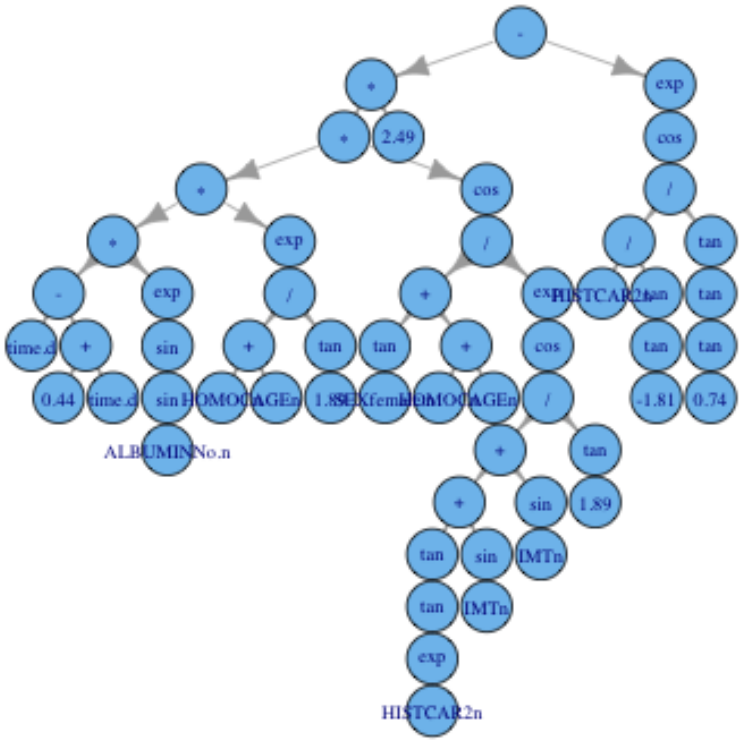


Figure 6.2: The final model developed by genetic programming, presented as a binary tree.

peated 25 times, using bootstrap re-sampling to better understand the frequencies at which different subsets of predictors were selected. In the pool of 25 backwards selected Cox models the mean number of predictors used was 9 (IQR: 8—10). There was a reasonable association between the estimated effect of a predictor according in the reduced backwards step-wise model and the frequency of the selection when the step-wise selection was repeated in the bootstrap procedure (table 6.8).

Generally the features selected by repeating the GP were far more variable than the

Table 6.8: Number (proportion) of times predictors were selected duting the 25 repititions of Cox regression backwards step-wise selection procedure and genetic programming.

Predictor		Cox Regression	Genetic Programming
Age	AGE	25 (1.00)	19 (0.76)
Gender	SEX	11 (0.44)	5 (0.20)
Smoking	SMOKING	5 (0.20)	10 (0.40)
Alcohol	ALCOHOL	3 (0.12)	11 (0.44)
Body mass index	BMI	19 (0.76)	4 (0.16)
Systolic, by hand	SYSTH	14 (0.56)	6 (0.24)
High-density lipoprotein cholesterol	HDL	18 (0.72)	8 (0.32)
Diabetes	DIABETES	10 (0.40)	1 (0.04)
Previous atherosclerosis (sum score)	HISTCAR2	25 (1.00)	20 (0.80)
Homocysteine	HOMOC	19 (0.76)	13 (0.52)
Creatinine	CREAT	22 (0.88)	7 (0.28)
Albumin	ALBUMIN	22 (0.88)	23 (0.92)
Presence of carotid artery stenosis	STENOSIS	10 (0.40)	10 (0.40)
Intima media thickness	IMT	23 (0.92)	12 (0.48)

features selected by the Cox regression stepwise selection procedure (table 6.8). This is to be be expected as GP is a stochastic system, where as the stepwise procedure is deterministic, only giving variable results because we are repeating the procedure on different bootstrap resamples of the derivation data set.

Despite this, the predictors that were estimated to have the largest effect in the final stepwise selected Cox model which were also selected with the high frequency in bootstrapped stepwise selection—age, previous atherosclerosis, and homocysteine—were selected (relatively) frequently when GP was repeated. Interestingly, stepwise selection also often selected BMI, HDL, creatinine, albumin and, IMT as predictors, however,

these featured in a relatively low proportion of the GP models. However, albumin and IMT did feature in the best performing 'final' GP model. Conversely, the 'final' GP model featured gender as a predictor, a predictor that was estimated to have small effect and was selected in low proportion of Cox models by the stepwise selection procedure.

6.3.3 Model Validation

Using the validation data set, the average performance of the 25 'best of run' prediction models automatically generated by GP was compared with the calibrated final Cox model. Graphical comparisons of the $S(t)$ values produced by each model with those obtained by the KM method in the validation set are shown in figure 6.3. As can be seen from this figure, both the Cox and GP models produced similar values that had good agreement with the KM estimates in the earlier years. However, this agreement deteriorated in the latter years, where the KM estimates have high variability, as indicates by the large error bars. This high variation my be explained by the fact that with a median follow-up time of 3.3 years, there are far fewer events and number of subjects in the latter time periods. Whilst agreement deteriorated in the latter time-points, both models had generally acceptable overall agreement.

The discriminative performance in the validation set, according to the C-statistic, of the models at different time points is shown in table 6.9 and figure 6.4. From the C-statistic estimates we can see that a satisfactory performance of >0.6 was reached in both models, at all time points. There was generally comparable discriminative performance of both models, at all time points, albeit in favour of the Cox model. Both models demonstrated better performance at time $t = 3$ years, which may be explained by the 3.3 median follow up time in the validation set.

The calibration plots evaluated by grouping subjects according to quantiles of predicted risk ($1 - S(t)$) at t = 1, 3, and 5 years are shown in figure 6.5. The corresponding χ^2 statistics and p-values are shown in table 6.10. From the graphical inspection of

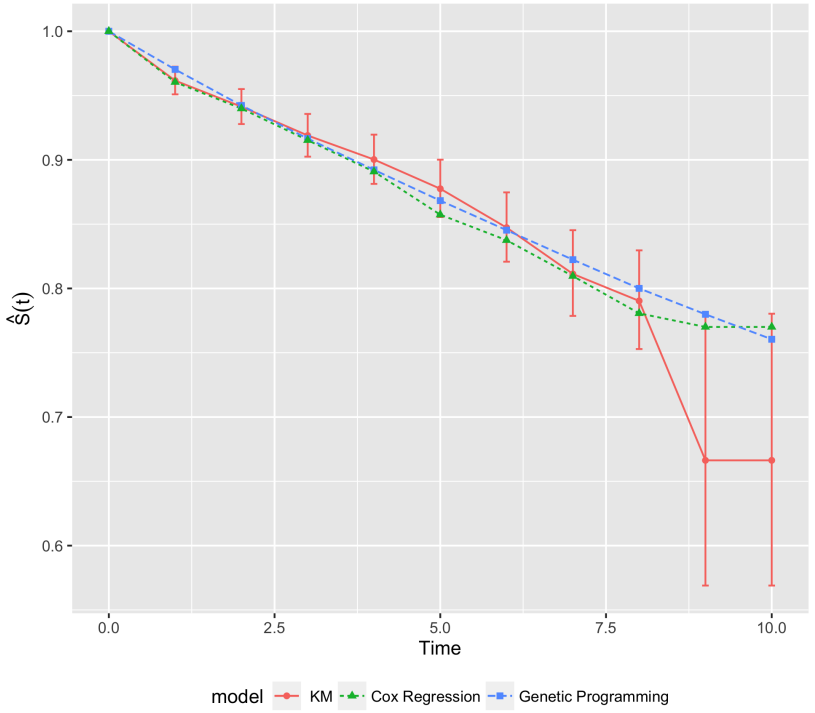


Figure 6.3: Average survival curves for the Cox regression and genetic programming models. The error bars represent ± 2 standard errors of the KM estimates.

Table 6.9: C-statistic estimates by model at t=1, 3, and 5 years

Time (years)	Cox PH Regression	Genetic Programming
1	0.66	0.59
3	0.70	0.69
5	0.70	0.64

the calibration plots we can see that there was no tendency to systematically over- or under-predict at any of the time points in either the Cox or GP models. The genetic programming model was less calibrated than the Cox model, confirmed by the higher χ^2 values in table 6.10 at times $t = 3$ and $t = 5$, whereas it was better calibrated at time $t = 1$. Calibration in both models was worst at time $t = 5$, and best in the Cox and GP

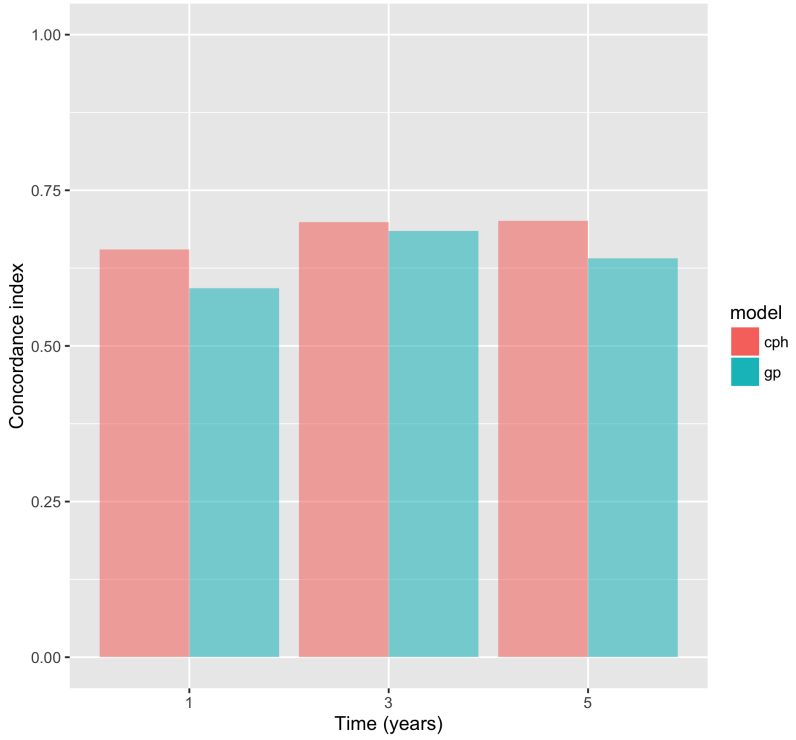


Figure 6.4: C-statistic estimates by model for t=1, 3 and 5 years

Table 6.10: χ^2 statistic for the comparison between observed versus expected (according to the model) number of events in groups of patients defined according to the predicted $1 - S(t)$ at t=1, 3, and 5 years.

Time (years)	Cox Regression		Genetic Programming	
t	χ^2	p-value	χ^2	p-value
1	7.93	0.541	5.18	0.818
3	4.89	0.844	9.99	0.352
5	10.32	0.325	16.17	0.063

models at times $t = 3$ and $t = 1$, respectively. However the Homser-Lemeshow test statistic, detailed in table 6.10, suggested that there was only a statistically significant lack of calibration in the GP model at time point $t = 5$.

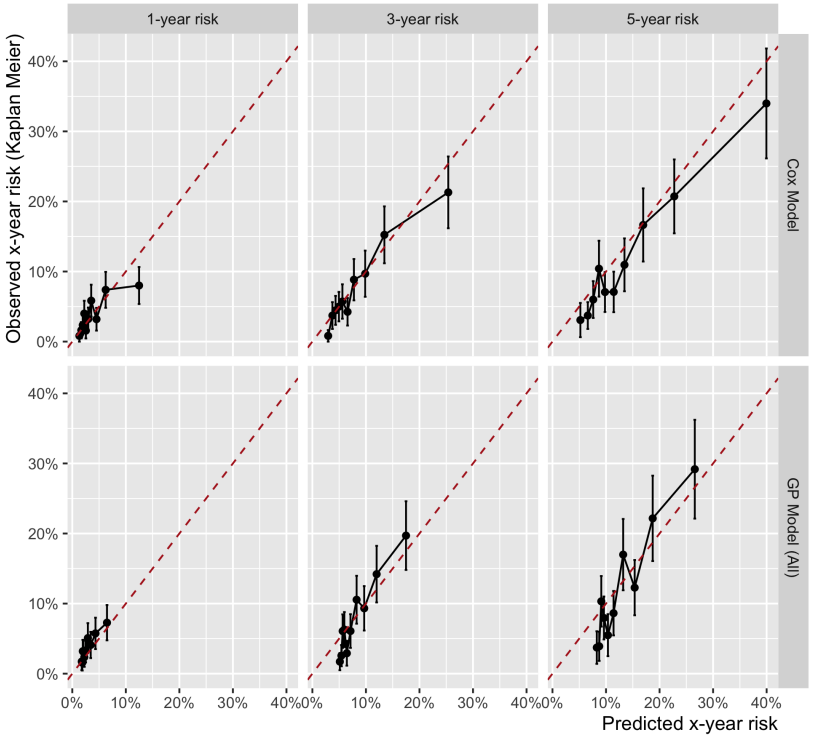


Figure 6.5: Calibration plots for the Cox regression and genetic programming models, at t=1, 3, and 5 years..

6.4 Discussion

This study showed that Cox regression and GP produced similar results when evaluated in a common validation data set. After re-calibration the discriminative ability of the GP model in the validation set was slightly larger than that of the Cox model at two time points, compared with the Cox model model, which was marginally better at only one time point. Despite slight relative differences, both models demonstrated an acceptable level of discriminative ability (C-index >0.6) at all times points. The GP model had relatively poorer calibration when compared with the Cox model. The Cox model demonstrated no statistically significant lack of calibration at any time point, however the GP did demonstrate a statistically significant lack of calibration at the latter time point only.

Despite generally comparable performance, albeit in favour of the Cox model, the predictors selected for representing their relationship with the outcome were quite different. The final reduced Cox model used 9 predictors, in contrast to 6 predictors used in the GP. The GP model used significantly fewer predictors, further confirmed by repeating the the GP and the stepwise selection procedure used in the Cox modelling, resulting in mean numbers of predictors of 6 (IQR: 5—8) and 9 (IQR: 8—10), respectively.

Predictors that were estimated to have larger contributions to the final Cox model and that frequently selected during stepwise selection—age, previous atherosclerosis, and homocysteine—were selected in the final GP model, with age and previous atherosclerosis selected at (relatively) high frequency when the GP was repeated. However, others predictors that had large to moderate contributions to the final Cox model—BMI, creatinine, and HDL—did not features in the final GP model and were selected infrequently when GP was repeated. Interestingly, gender did not have much of contribution to the final cox model yet it was selected relatively infrequently by GP and featured in the final GP model. There were other predictors —albumin and IMT —where the picture was less clear, and whilst they had large to moderate contributions to the final Cox model and featured in the final GP model, they were selected at low frequencies when GP was repeated. Whilst these results confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling, symbolic regression model did not estimate such a large number predictors to be strongly associated with the outcome, associated strongly enough for inclusion in the model at least, whilst achieving comparable performance.

These results suggest that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors. To test the first part of this hypothesis—that GP can represent the potential non-linear relationships that exist between predictors—the shape of the predictor effects were plotted to evaluate whether or not the effects were non-linear in nature. Using the 'final' GP model, the effects of each

predictor's values were plotted against log hazard, whilst the other values were held at their reference values. Reference values were the modal class for binary variables and the mean of continuous variables. Figure D.1 of appendix D illustrates that the the 'final' model developed by GP is modelling non-linear effects for the continuous predictors age, initma media thickness and homocysteine.

To test the second part of this hypothesis—that GP can better represent these relationships between predictors using fewer variables—we repeated the GP runs with exactly the same experimental set-up, but restricting the inputs to predictors that were selected with a relatively high frequency (>0.5) in the original GP run of the primary experiment. The covariates included age, previous atherosclerosis, homocysteine, and albumin. This produced very similar results (detailed in appendix E) both in terms of calibration and discrimination. Both these findings support the hypothesis that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

Whilst considerable effort was made to relax the linearity of the Cox regression, through transformation of predictors, the nature of the approach relies on linear combinations of predictors. The fact that GP required fewer predictors to achieve similar performance may have an advantage in practical application of the developed clinical prediction model. The acquisition of information that forms the inputs to such a model can be prohibitively onerous in routine clinical practice. Therefore a prediction model that requires fewer inputs, especially if the information relating to these inputs is in practice recorded easily and to a good quality, would considerably increase adoption and utility.

This work has limitations introduced by its use of data from the SMART study, a study from a secondary prevention setting, designed to predict the risk of subsequent cardiovascular events in patients with already presenting with clinical CVD. Through its use of the SMART study data this work has demonstrated the utility of GP in a secondary prevention setting, however, there are limitations in the generalisability of these findings to the other clinical settings of cardiovascular risk prediction. Indeed,

secondary prevention in stable cardiovascular patients is not the most common clinical setting for the application of cardiovascular risk prediction models in routine practice.

6.5 Conclusion

To our knowledge, this is the first empirical study to assess the value of GP for clinical prediction purposes compared to the well-known and widely applied Cox PH regression technique. Using data from the SMART study we demonstrated that a symbolic regression model developed by SSOGP has predictive ability comparable to that of Cox regression for the prediction of future cardiovascular events in a secondary prevention setting, i.e in patients with symptomatic cardiovascular disease. These experiments compared an untuned SSOGP symbolic regression model that was developed in an automated fashion using basic parameter values recommended from the GP literature, with a highly tuned Cox regression model that was developed in a very involved manner that required a certain amount of clinical and statistical expertise.

Whilst the highly tuned Cox regression model performed marginally better in the validation data, both in terms of calibration and discrimination, the performance of the automatically generated prediction model was generally comparable.

These findings demonstrate the utility of GP as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes, where the primary goal is accurate prediction. These findings also confirm the prognostic significance of age, previous atherosclerosis, homocysteine, and to a lesser extent albumin and IMT, for cardiovascular risk in patients with symptomatic CVD. Finally, further validation is required to assess the utility of GP for automated development of new clinical prediction models in other clinical and environmental settings.

In the next chapter we discuss a series of experiments using a very similar methodology and experimental set-up to those used in this chapter. However, the developed

GP approach is applied to different dataset refined from CPRD in a primary prevention setting in patients with T2DM, rather than the secondary prevention setting in the general population used in this chapter. We look to see if we observe similar results to this chapter and further demonstrate the utility of GP for clinical prediction modelling in censored survival data in different clinical setting.

Chapter 7

Experiment 3: A case study in asymptomatic cardiovascular disease in type 2 diabetes using CPRD

Next we perform our third and final set of experiments—which have a very similar experimental set-up those in the previous chapter—to further test our main hypothesis that application of GP can provide more accurate representation of factors that predict the risk of CVD when compared with existing methods, but this time using a different clinical setting and datasource.

We assess the utility of the developed SSOGP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events, assessing its performance and examining the prognostic significance of different risk factors when compared with the *de facto* statical method in a much larger observational cohort of patients from CPRD in a primary prevention clinical setting, where patients have asymptomatic CVD.

Background & Aims The aim of this study was to demonstrate the utility of genetic programming for the automatic development of clinical prediction models using symptomatic cardiovascular disease in patients with T2DM as a case study.

Study Design and Setting We compared genetic programming against the commonly used Cox regression technique in terms of development and performance of a cardi-

ovascular risk score using data from CPRD to refine a retrospective observations cohort of T2DM patients with asymptomatic cardiovascular disease. The event predicted was a composite cardiovascular event, comprising of cardiovascular death, non-fatal stroke, and myocardial infarction. The predictive ability of both models was assessed in terms of discrimination and calibration.

Results The study cohort consisted of 63,496 patients with T2DM, registered with practices between 1999 and 2011, aged 35-85 years with a total of 14,804 cardiovascular events. The study cohort was split 70:30 into derivation and validation sets, used for model fitting and assessment of performance of both the genetic programming and Cox regression models. The discrimination of both models was comparable, albeit in favour in Cox regression; at time points t=2, 5, and 8 years the C-index was 0.69, 0.65, 0.67, and 0.71, 0.70, 0.70, for the genetic programming and Cox regression models, respectively. At the same time points, the calibration of both models was also comparable, with no significant lack calibrated to the validation data.

Conclusions Using empirical data, we have confirmed the findings of the previous chapter in a new clinical context—primary prevention—demonstrating that a prediction model developed automatically by genetic programming has predictive ability comparable to that of manually 'tuned' Cox regression. The genetic programming model was more complex but was developed in a fully automated fashion, used significantly fewer predictors as inputs, and did not require the expertise needed for survival analysis. Genetic programming demonstrated potential as a methodology for the automated development of clinical prediction models for diagnostic and prognostic purposes.

7.1 Introduction

CVD is the leading cause of mortality and a major cause of morbidity globally and in the UK. Asymptomatic patients that are suspected to be at high risk need to be identi-

fied by general practitioners so they can offer advice about lifestyle changes and initiate preventative treatment. To facilitate this, general practitioners need tools that can accurately and reliably predict cardiovascular risk in their patients. National policies for the management of both CVD and type 2 diabetes advocate the calculation of CVD risk in order to identify high-risk patients for targeted interventions [264, 295, 245, 66, 216]. As discussed in chapter 3, several multivariable risk-prediction models have been developed for the general, non-diabetic population that also account for diabetes, but only a few are specific to type 2 diabetes [303].

In chapter 5 we carried out an external validation of the performance of the UKPDS-RE on a large, relatively contemporary retrospective cohort of UK-resident patients with T2DM from CPRD. Results showed that the UKPDS-RE had a reasonable ability to identify high-risk patients (discrimination) but were generally poor at quantifying the absolute risk (calibration). Our findings suggested that the use of UKPDS-RE in clinical practice will lead to over-estimation of CVD risk in patients with newly diagnosed T2DM. Considering the widespread application of these prediction models in clinical practice, drug reimbursement, and public health decision-making, we suggest that there is a need for revised risk equations in T2DM.

The objective of this study was to compare the GP approach for censored longitudinal data developed in section 6.2.1, with multi-variable Cox regression in the development of a clinical prediction model for the occurrence of vascular events in a large, relatively contemporary dataset of UK-resident patients with T2DM. Models were developed using SSOGP and multi-variable Cox regression, and their performance was evaluated in terms of discrimination and calibration in a validation data set.

The experiments in this chapter differ from chapter 6 in terms of the clinical setting and the cohort of patients. In the previous experiment we assessed the developed GP approach in the clinical setting of secondary prevention, where we are predicting subsequent cardiovascular events in patients with a clinical diagnosis of CVD, using data from a prospective cohort study designed to identify predictors of future cardiovascular

events in patients with symptomatic CVD. In this chapter we evaluate the developed GP approach in the primary prevention setting, where we are predicting the risk of a primary cardiovascular event in a much larger retrospective observational cohort of patients with T2DM from UK general practice.

7.2 Patients and Methods

This study was carried out using data from CPRD and linked data from the ONS and HES. Ethical approval for the study was granted by the CPRD Independent Scientific Advisory Committee on 9th October 2012, protocol number 12_111R (Appendix F). The CPRD dataset and its associated linked datasets have been described previously in sections 2.6 and .

This study considers a prospective open cohort of CVD-free patients with T2DM in the January 2013 build of CPRD, over 14 years from 1997 to 2011, aged 30-85 years at index date and registered with practices participating in CPRD-HES/ONS linkage to ensure accurate cause of death, ethnicity and socioeconomic status. For each patient, the start of follow-up is defined as the latest of: patient registration date, practice Up-To-Standard (UTS) date, start of HES coverage, and start of ONS coverage; the index date is then calculated as the start of follow-up plus 365 days wash-in (see below). End of follow-up is defined as the earliest of: patient transfer-out date (where not internal), ONS death date, practice last-data-collection date, patient HES linkage date, end of HES coverage, and end of ONS coverage. In addition, the ONS SES linkage period must overlap the follow-up period by at least one day.

Patients were excluded from the cohort that have any one of the following criteria: a recorded diagnosis of cardiovascular or cerebrovascular disease prior to the index date; any temporary residence status; interrupted periods of registration with the practice; no valid Index of Multiple Deprivation (SES); were taking statins at index; implausible or improbable dates; or recorded risk factor values out of plausible range. Patients

were selected that were eligible for linkage schemes with the HES, ONS mortality data and Index of Multiple Deprivation data throughout their respective period of follow-up. This should provide accurate ascertainment of ethnicity, socioeconomic status, and cause of death. Issues with ethnicity data where within the non-missing data there are a large proportion of ethnicities recorded as 'unknown' will be addressed by recoding the 'unknown' responses as 'white', with the rationale that, assuming the study population is comparable with the UK population, 93% or more of people without ethnicity recorded would be expected to be from a white ethnic group.

Patients were considered for selection if they had a clinical (Read or ICD-10) code indicative of diabetes mellitus in their CPRD or linked HES records. As not all clinical codes for diabetes distinguish between type 1 diabetes and type 2 diabetes, and some patient histories may erroneously have contained both type 1 and type 2 diabetes codes, these patients were categorised as having type 2 diabetes if they met one or more of the following criteria:

- Clinical codes exclusively indicative of type 2 diabetes
- At least one clinical code indicative of type 2 diabetes (regardless of others indicative of type 1 or non-specific diabetes) and at least one prescription for an Oral Hypoglycaemic Agent (OHA)
- Prescription of two or more classes of OHA
- Diagnoses of both type 1 and type 2 diabetes and an age of diagnosis older than 35 years.

Any patient with evidence of diabetes secondary to other causes was excluded. The date of diabetes incidence was defined as the date of either first diagnosis or first prescription of a diabetes medication, whichever was earlier. A 'wash-in' period of 365 days was applied to exclude non-incident T2DM cases.

The primary outcome is first CVD event either before or at death, with CVD being defined here as coronary heart disease (including myocardial infarction and angina) or cerebrovascular disease (including stroke and transient ischemic attack). A wash-in of 365 days is applied to ensure no prior history of CVD. Either a diagnosis of CVD or an intervention to treat CVD—such as an angioplasty or coronary bypass—will be considered an event. These events will be recorded as Read codes in the CPRD Clinical or Referral tables; as ICD-10 or OPCS codes in HES diagnosis or procedure tables, respectively; or as ICD-10 or ICD-9 codes in the OPCS cause-of-death data. Lists of the relevant codes are supplied as part of the approved ISAC protocol for this study (Appendix F).

For our study we *a priori* selected 23 candidate predictors—values for which were taken from CPRD observations around the index date—based on previous prognostic studies (e.g UKPDS). These 23 candidate predictors included indicators of baseline comorbidity (Charleston Index, no. general practitice attendances in year prior, treated hypertension, durations of diabetes, a recorded diagnosis of the following; some other form of CVD not defined in the outcome, renal disease, rheumatoid arthritis, atrial Fibrillation), demographics (age, gender, self assigned ethnicity, and quintiles of the Index of Multiple Deprivation) and risk factors for vascular events in the general population (smoking status, BMI, SBP, and lipids). Treated hypertension is defined as diagnosis of hypertension and at least one current prescription of at least one antihypertensive agent(e.g. thiazide, Beta-blocker, calcium channel blocker, or Angiotensin-Converting-Enzyme (ACE) inhibitor). Prescribed related drugs (lipid-lowering, ACE, Angiotensin Receptor Blockers (ARB), Beta-blocker, and Anti-Platelet Therapy (APT) therapies), table 6.5) were also considered as it is conceivable that they are relevant to predict future events in patients with asymptomatic vascular disease. We note that the primary focus of these models is achieving accurate predictions rather than insight into the predictor effects.

Multiple imputation was considered to replace missing values for ethnicity, smoking

status, BMI, SBP, TC, HDL, LDL and triglycerides. However, the proportions of missing data exceeded what can be reliably imputed using even the more advanced multiple imputation techniques. Instead categorical predictors were given an additional 'missing' categories and continuous predictors were categorised into clinical meaningful categories, also with an additional 'missing' category. In the clinical context of primary prevention thus level of missing data for many predictors in question is not unexpected, as they would not normally be recorded unless the general practitioner already suspects some above average risk of CVD. Whilst, to some degree, in categorising the continuous predictors that have missing values we discard some of the information, this loss is outweighed by the loss of not including these predictors at all, or in removing observations with these missing values (complete case analysis) as this would considerably reduce the sample size and in turn the power to detect the patterns in the data.

7.2.1 Methods

The data set was split, randomly, into two parts: a derivation set of approximately 66.67% (42,331 patients) and a validation set of approximately 33.33% (21,165 patients). The derivation set was used for model development (both by Cox regression and by genetic programming) and the validation set to access the performance of the two models. The aim for both models was to predict the absolute risk of occurrence of vascular events (stroke, myocardial infarction or cardiovascular death). Given the available follow-up, 2-, 5-, and 8-year risks could be assessed. With respect to sample size in the derivation set, the balance of 9,878 events and 23 predictors is reasonable, (table 6.5). At least 10-20 events per candidate predictor have been proposed in previous guidelines for the sensible development of predictions models [112, 236, 285, 286].

Cox Regression

In the derivation set, we fitted a Cox regression model using a similar modelling strategy that was described in chapter 6 in the development of a clinical prediction model in the cohort developed from the CPRD data set. Briefly, the we first fitted a full main effects model. Biologically implausible values were set to missing (prior to imputation) and extreme values truncated at the 1st and 99th centile. To enhance the flexibility of the Cox regression and enable fairer comparison with the (unrestricted) genetic programming, we considered continuous predictors (e.g. age, duration of diabetes) for transformation. Several transformations were considered in adding polynomials, fractional polynomial terms, transformations (e.g. log, square root, exponential), restricted cubic splines (with varying number of knots) and linear coding (i.e categorisation). To further enhance a fair comparison with genetic programming, we considered interaction effects between predictors.

Key limitations of the Cox PH model include the assumption of proportional hazards - that hazard functions in the different strata are proportional over time, assumptions of linearity and additivity which are implicit in regression's linear combinations, and the fact that the baseline hazard is never specified (although this last one may be advantage in some cases). All model assumptions relevant to the Cox proportion hazards model were tested. A reduced model was obtained by applying a backwards selection procedure, with Akaike information criterion (AIC) as the stopping criterion.

Internal validation of the model was performed using a bootstrapping re-sampling procedure [22, 75, 111]. Random samples were drawn (with replacement) from the derivation set with 200 replications, and the backwards selection of predictors for the reduced model repeated each time. Bootstrapping yielded an estimate of optimism of the reduced models as expressed by the concordance (C) statistic, which for a binary outcome is identical to the area under the receiver operating characteristic (ROC) curve. A shrinkage factor was derived from the bootstrap estimates to re-calibrate the model to adjust for optimism. The re-calibrated model was applied to the validation set

to estimate its discrimination and calibration in an independent sample. All analyses were carried out in R (v3.0.1) [247].

Symbolic Regression

For the experiments in this chapter we implemented the same tree-based SSOGP approached as in the previous SMART experiments chapter (discussed in chapters 4 and 6) to fit symbolic regression models to the data to estimate discrete hazard, thus predicting the risk for cardiovascular events. Findings from the previous experiments in the SMART study data (chapter 6) found that the GSOGP and, GMOGP search strategies were too computationally expensive for the computing resources available. As the CPRD cohort used in these experiments is considerably larger than that of the SMART cohort, and thus more computationally expensive, the generational GP approached were not considered in this experiment.

Symbolic regression was performed using the RGP package in the R statistical programming language, with the SSOGP search heuristic using the same parameter settings as the previous experiments, detailed in table 7.1.

Briefly, the default RGP function set ($+, -, \div, \times, \sin, \cos, \tan, \sqrt{}, \exp, \log$) was used to enable the representation of potential non-linear relationships present in the training data. Koza's [167] ramped half-and-half random initialisation method was used with a maximum tree depth of 156, which was calculated as a function of the expected solution depth of 52. The GP approach utilised is untyped with the search space constrained by the use of fitness penalties. Specifically, the 'death penalty' is used where invalid solutions, such invalid mathematical operations (e.g. dividing by 0), have their fitness value set to ∞ giving them the lowest possible fitness (section 4.2.1).

Again, as in the previous experiments, expiration of a fixed time compute budget was chosen as the termination criterion. The compute time budget for these experiments was set to 12 hours wall-time and these experiments were run on a single thread on an

Intel Westmere 2.8GHz CPU with 48 GB of memory.

Table 7.1: Parameters of the SSOGP search heuristic.

	Variable (Symbol)	Domain	Setting
Population Size	mu (μ)	$\mathbb{N}$	1,000
Tournament Size	tournamentSize ($s_{tournament}$)	$\mathbb{N}$	4
Recombination Probability	recombinationProbabilty (p_{rec})	$[0, 1]$	0.5

We did not perform internal validation in the genetic programming approach using a bootstrap as we did with the Cox regression, because it would not have been possible to convert it into a shrinkage factor in the same way as we would for a Cox regression. The genetic programming system is a stochastic process, with each run potentially yielding models with differing complex structures (i.e. symbolic regression). As a result regression coefficients do not exist in genetic programming models in the same way that they do in regression models. Instead the training data was split 2/3:1/3 into training and holdout sets, using a stratified random split to ensure proportionate number of events. The first 2/3, the training set, was used for training to induce a population of prediction models. The remaining 1/3, the holdout set, was used at the end of the genetic programming run to calculate the fitness of the population of models and thus determine the fittest or 'best of run' model to be returned as the output of the genetic programming system. In this way the final genetic programming model was selected based on its fit to unseen data using a sample other than which it was trained or developed.

To understand variable selection in the genetic programming and enable comparison with bootstrapped backwards selection of the Cox model, the genetic programming system was executed 25 times to produce 25 'suggested' models. For each iteration the training data was randomly (stratified) split 2:1. The final genetic programming model was applied to the validation data set to assess its performance, in terms of discrimination and calibration in an independent sample. All analyses were carried out

in R version 3.1.2 [247].

Comparison of both methods

The two clinical prediction models, one obtained from Cox progression and three from symbolic regression (SSOGP), were evaluated in terms of overall survival curves, discrimination and calibration in the validation data set. The models were used to predict the discrete hazards $h(t)$ at t = 2, 5, and 8 years. Model were first evaluated visually by comparing the survival probabilities $S(t)$ predicted by the models with estimates obtained using the KM method. The agreement between these curves and the KM estimates were assessed visually.

Discrimination was assessed using the concordance statistic (C-statistic) [109, 110, 111, 300], which was evaluated considering truncation of the survival/censoring times at t=2, 5, and 8 years.

Model calibration was assessed using calibration plots and the generalisation of the Hosmer-Lemeshow test statistic for survival data [56]. Calibration was evaluated by grouping subjects according to the predicted $S(t)$ at t = 2, 5, and 8 years. All analyses were carried out in R version 3.1.2.

7.3 Results

7.3.1 Descriptives

There were no major differences in the baseline characteristics of the patients between the derivation and validation sets (table 7.2). Data were available on 255,478 and 126,937 person-years collected during a median follow-up of 5.33 (range, 0-13 years) and 5.25 years (0-13 years) for the derivation and validation sets, respectively. In

the derivation set a total of 9,834 events occurred, corresponding to 2-, 5-, and 8-year cumulative incidences of 6.8%, 17.0% and 26.0% respectively. In the validation set a total of 4,970 events occurred, corresponding to 2-, 5-, and 8-year cumulative incidences of 6.8%, 17.0% and 27.0% respectively.

7.3.2 Model Derivation

Prior to modelling extreme values in continuous predictors were truncated at the 1st and 99th centile. Categorical predictors with missing values were given an additional 'missing' categories and continuous predictors with missing values were categorised into clinical meaningful categories, also with an additional 'missing' category.

The full Cox regression model consisted of 23 predictors, some of which had limited contributions. Predictors that had a relatively large effect were age, ethnicity, gender, smoking status, SES, no. of general practitioner contacts in previous year, recorded diagnosis of 'other' CVD, atrial fibrillation, BMI, SBP, lipids, ACE/ARB therapy, and APT therapy. The proportionality of hazards was tested using an overall test which was not significant. We judged our sample size to large enough to allow for some model reduction (9,834 events and a full model with 39 degrees of freedom), facilitating easier practical application and clinical interpretation. We applied a backwards step-wise selection procedure, using AIC as the stopping rule, to achieve a reduced Cox model. The reduced step-wise selected model was found to be optimal with 20 predictors (table 7.3). Predictors with relatively weaker effects (treated hypertension, LDL, and duration of T2DM) were excluded from the reduced model.

Bootstrapping of the reduced model yielded an estimate of required shrinkage for the coefficients in the step-wise selected model of 0.99, suggesting that each coefficient should be reduced by 1% to obtain a re-calibrated model that corrects for optimism. This shrinkage factor was applied to the reduced backwards step-wise model and considered the calibrated 'final' Cox regression model (table 7.4).

Table 7.2: Baseline characteristics of patients in the CPRD cohort, by derivation and validation sets (n=63,496).

		N	Test set	Training set	Test Statistic
			$N = 21165$	$N = 42331$	
Cardiovascular event		63496	23% (4970)	23% (9834)	$\chi^2_1 = 0.5,\ P = 0.48^1$
Age (at baseline)	years	63496	52 62 71	52 62 71	$F_{1,63494} = 0.19,\ P = 0.66^2$
Ethnicity : Mssn		63496	18% (3816)	18% (7673)	$\chi^2_3 = 0.44,\ P = 0.93^1$
Nn-W			8% (1638)	8% (3219)	
Unkn			15% (3174)	15% (6371)	
Whit			59% (12537)	59% (25068)	
Gender : Feml		63496	47% (9909)	47% (19736)	$\chi^2_1 = 0.22,\ P = 0.64^1$
Smoking status : Crrn		63496	19% (3940)	19% (7999)	$\chi^2_3 = 1.4,\ P = 0.71^1$
Frmr			23% (4837)	23% (9754)	
Nevr			51% (10738)	50% (21290)	
Mssn			8% (1650)	8% (3288)	
Index of Multiple Deprivation	quintiles : 1	63496	18% (3741)	18% (7804)	$\chi^2_4 = 6.2,\ P = 0.19^1$
2			22% (4695)	22% (9197)	
3			20% (4315)	20% (8603)	
4			21% (4456)	21% (8898)	
5			19% (3958)	18% (7829)	
Charlson index : 0		63496	3% (560)	3% (1130)	$\chi^2_4 = 4.8,\ P = 0.31^1$
1			66% (13900)	65% (27595)	
2			19% (3979)	19% (8254)	
3			9% (1800)	8% (3540)	
4			4% (926)	4% (1812)	
No. GP attendances year prior		63496	5 9 15	5 9 15	$F_{1,63494} = 0.04,\ P = 0.85^2$
Recorded diagnosis of Other CVD		63496	17% (3695)	18% (7499)	$\chi^2_1 = 0.64,\ P = 0.42^1$
Treated hypertension		63496	39% (8280)	39% (16316)	$\chi^2_1 = 2,\ P = 0.16^1$
Renal disease		63496	7% (1498)	7% (3028)	$\chi^2_1 = 0.12,\ P = 0.73^1$
Rheumatoid arthritis		63496	1% (297)	2% (652)	$\chi^2_1 = 1.8,\ P = 0.18^1$
Atrial Fibrillation		63496	3% (615)	3% (1301)	$\chi^2_1 = 1.4,\ P = 0.24^1$
Duration of T2DM	days	63496	17 49 106	17 49 106	$F_{1,63494} = 0.02,\ P = 0.9^2$
Body mass index	Kg/m2	45656	26 29 33	26 29 33	$F_{1,45654} = 0.79,\ P = 0.37^2$
Systolic blood pressure	mm Hg	51930	130 140 150	130 140 150	$F_{1,51928} = 0,\ P = 0.95^2$
Total cholesterol	mmol/L	37817	4.2 4.9 5.7	4.2 4.9 5.7	$F_{1,37815} = 0.5,\ P = 0.48^2$
High-density lipoprotein cholesterol	mmol/L	23071	1.0 1.2 1.4	1.0 1.2 1.4	$F_{1,23069} = 2.5,\ P = 0.11^2$
Low-density lipoprotein cholesterol	mmol/L	18470	2.1 2.7 3.4	2.1 2.6 3.4	$F_{1,18468} = 0.49,\ P = 0.48^2$
Triglycerides	mmol/L	27390	1.2 1.7 2.5	1.2 1.7 2.5	$F_{1,27388} = 0.46,\ P = 0.5^2$
Lipid lowering Tx		63496	31% (6624)	31% (13224)	$\chi^2_1 = 0.02,\ P = 0.88^1$
ACE/ARB Tx		63496	35% (7349)	34% (14596)	$\chi^2_1 = 0.36,\ P = 0.55^1$
Beta-blockers Tx		63496	14% (2902)	13% (5401)	$\chi^2_1 = 11,\ P < 0.001^1$
Anti-Platelet Tx		63496	20% (4262)	20% (8516)	$\chi^2_1 = 0,\ P = 0.95^1$

$a\ b\ c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. N is the number of non–missing values. Numbers after percents are frequencies. NA represent missing values. Tests used:[1]Pearson test; [2]Wilcoxon test. Tx represents therapy.

Table 7.3: Cox regression coefficients in the full model, and stepwise selected model (using AIC).

Predictor		Full	Stepwise
ACE/ARB Tx	ACEARB=1	0.1008	0.1088
Atrial Fibrillation	AF=1	0.2032	0.1992
Age (at baseline)	AGE	0.0445	0.0441
Anti-Platelet Tx	APT=1	0.2634	0.2605
Beta-blockers Tx	BETAB=1	0.0811	0.0857
Body mass index	BMIf=O/P-	−0.1166	−0.1174
	BMIf=Obes	−0.0613	−0.0630
	BMIf=Un/N	−0.0528	−0.0547
Charlson index	CHARLS	0.0346	0.0342
No. GP attendances year prior	CONT	0.0106	0.0106
Recorded diagnosis of Other CVD	CVD.other=1	0.2548	0.2516
Ethnicity	ETHNICf=Nn-W	1.4759	1.4617
	ETHNICf=Unkn	1.2071	1.1957
	ETHNICf=Whit	1.3145	1.3031
High-density lipoprotein cholesterol	HDLf=Best	−0.0675	−0.0664
	HDLf=Bttr	−0.1135	−0.1110
	HDLf=Mssn	0.0454	0.0715
Treated hypertension	HYPER=1	0.0164	
Low-density lipoprotein cholesterol	LDLf=Mssn	0.0269	
	LDLf=NrOp	0.0355	
	LDLf=Optm	−0.0811	
Lipid lowering Tx	LLT=1	−0.0449	−0.0512
Rheumatoid arthritis	RA=1	0.1194	0.1195
Renal disease	RENAL=1	0.1069	0.1041
Systolic blood pressure	SBPf=Hy/N	−0.0739	−0.0767
	SBPf=Mssn	0.0077	0.0070
	SBPf=NrmH	−0.0805	−0.0825
Index of Multiple Deprivation	SES_5	0.0533	0.0528
Gender	SEX=Feml	−0.2142	−0.2124
Smoking status	SMOKf=Frmr	−0.2166	−0.2134
	SMOKf=Mssn	−0.1548	−0.1506
	SMOKf=Nevr	−0.2705	−0.2670
Duration of T2DM	T2DM.dur	0.0000	
Total cholesterol	TCf=BrdH	0.1012	0.1162
	TCf=High	0.1870	0.1996
	TCf=Mssn	0.0862	0.0909
Triglycerides	TRIGf=Dsrb	−0.0736	−0.0760
	TRIGf=High	0.0631	0.0643
	TRIGf=Mssn	0.0254	0.0328

Table 7.4: Association of each predictor with cardiovascular events in the calibrated final Cox model.

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
AGE	52	71	19	0.8378000	0.019885	0.7988200	0.8767700
Hazard Ratio	52	71	19	2.3113000		2.2229000	2.4031000
SES_5	2	4	2	0.1055100	0.015017	0.0760730	0.1349400
Hazard Ratio	2	4	2	1.1113000		1.0790000	1.1445000
CHARLS	1	2	1	0.0342070	0.012211	0.0102730	0.0581410
Hazard Ratio	1	2	1	1.0348000		1.0103000	1.0599000
CONT	5	15	10	0.1058000	0.012740	0.0808310	0.1307700
Hazard Ratio	5	15	10	1.1116000		1.0842000	1.1397000
ETHNICf — Mssn:Whit	4	1		-1.3031000	0.053376	-1.4077000	-1.1985000
Hazard Ratio	4	1		0.2716900		0.2447000	0.3016500
ETHNICf — Nn—W:Whit	4	2		0.1585700	0.040632	0.0789350	0.2382100
Hazard Ratio	4	2		1.1718000		1.0821000	1.2690000
ETHNICf — Unkn:Whit	4	3		-0.1074100	0.028978	-0.1642000	-0.0506120
Hazard Ratio	4	3		0.8981600		0.8485700	0.9506500
SEX — Feml:Male	1	2		-0.2123800	0.021353	-0.2542400	-0.1705300
Hazard Ratio	1	2		0.8086500		0.7755100	0.8432200
SMOKf — Crrn:Nevr	3	1		0.2669500	0.028446	0.2112000	0.3227100
Hazard Ratio	3	1		1.3060000		1.2352000	1.3809000
SMOKf — Frmr:Nevr	3	2		0.0535600	0.026013	0.0025756	0.1045400
Hazard Ratio	3	2		1.0550000		1.0026000	1.1102000
SMOKf — Mssn:Nevr	3	4		0.1163900	0.037252	0.0433750	0.1894000
Hazard Ratio	3	4		1.1234000		1.0443000	1.2085000
CVD.other — 1:0	1	2		0.2516000	0.026515	0.1996300	0.3035700
Hazard Ratio	1	2		1.2861000		1.2210000	1.3547000
RENAL — 1:0	1	2		0.1040500	0.042783	0.0202000	0.1879100
Hazard Ratio	1	2		1.1097000		1.0204000	1.2067000
RA — 1:0	1	2		0.1195100	0.073803	-0.0251380	0.2641600
Hazard Ratio	1	2		1.1269000		0.9751700	1.3023000
AF — 1:0	1	2		0.1992400	0.049470	0.1022800	0.2962000
Hazard Ratio	1	2		1.2205000		1.1077000	1.3447000
BMIf — Mssn:Obes	2	1		0.0629870	0.033153	-0.0019920	0.1279700
Hazard Ratio	2	1		1.0650000		0.9980100	1.1365000
BMIf — O/P—:Obes	2	3		-0.0544540	0.029073	-0.1114400	0.0025290
Hazard Ratio	2	3		0.9470000		0.8945500	1.0025000
BMIf — Un/N:Obes	2	4		0.0082777	0.034289	-0.0589270	0.0754830
Hazard Ratio	2	4		1.0083000		0.9427800	1.0784000
SBPf — Hy/N:Hypr	1	2		-0.0766970	0.050911	-0.1764800	0.0230880
Hazard Ratio	1	2		0.9261700		0.8382100	1.0234000
SBPf — Mssn:Hypr	1	3		0.0070265	0.033982	-0.0595770	0.0736310
Hazard Ratio	1	3		1.0071000		0.9421600	1.0764000
SBPf — NrmH:Hypr	1	4		-0.0824550	0.026521	-0.1344400	-0.0304740
Hazard Ratio	1	4		0.9208500		0.8742100	0.9699900
TCf — Dsrb:Mssn	4	1		-0.0909320	0.034001	-0.1575700	-0.0242920
Hazard Ratio	4	1		0.9130800		0.8542100	0.9760000
TCf — BrdH:Mssn	4	2		0.0252240	0.034996	-0.0433660	0.0938140
Hazard Ratio	4	2		1.0255000		0.9575600	1.0984000
TCf — High:Mssn	4	3		0.1086300	0.039523	0.0311700	0.1861000
Hazard Ratio	4	3		1.1148000		1.0317000	1.2045000
HDLf — Poor:Mssn	4	1		-0.0715470	0.039955	-0.1498600	0.0067627
Hazard Ratio	4	1		0.9309500		0.8608300	1.0068000
HDLf — Bttr:Mssn	4	2		-0.1825500	0.059112	-0.2984000	-0.0666910
Hazard Ratio	4	2		0.8331500		0.7420000	0.9354800
HDLf — Best:Mssn	4	3		-0.1379300	0.052705	-0.2412300	-0.0346290
Hazard Ratio	4	3		0.8711600		0.7856600	0.9659600
TRIGf — BrdH:Mssn	4	1		-0.0327550	0.048791	-0.1283800	0.0628740
Hazard Ratio	4	1		0.9677800		0.8795200	1.0649000
TRIGf — Dsrb:Mssn	4	2		-0.1088000	0.041304	-0.1897600	-0.0278510
Hazard Ratio	4	2		0.8969100		0.8271600	0.9725300
TRIGf — High:Mssn	4	3		0.0315420	0.042669	-0.0520870	0.1151700
Hazard Ratio	4	3		1.0320000		0.9492500	1.1221000
LLT — 1:0	1	2		-0.0511850	0.028526	-0.1070900	0.0047251
Hazard Ratio	1	2		0.9501000		0.8984400	1.0047000
ACEARB — 1:0	1	2		0.1088200	0.023102	0.0635440	0.1541000
Hazard Ratio	1	2		1.1150000		1.0656000	1.1666000
BETAB — 1:0	1	2		0.0856990	0.028756	0.0293370	0.1420600
Hazard Ratio	1	2		1.0895000		1.0298000	1.1526000
APT — 1:0	1	2		0.2605300	0.026711	0.2081800	0.3128900
Hazard Ratio	1	2		1.2976000		1.2314000	1.3674000

Figure 7.1 describes the run statistics for the 25 SSOGP runs performed, based on the parameter settings detailed in section 7.2.1, on the different stratified re-samples of the derivation data set. The figure depicts the evolution of the different GP run's best fitness (bestFit) and complexity, quantified using mean visitation length (meanVisLen), over time. Time is represented as iterative or evolutionary steps (stepNo), where an evolutionary step is each time that tournament selection is performed and new individuals are generated and considered for inclusion in the population. The 'final' symbolic regression model is the individual the best (i.e. lowest) fitness at the end of all 25 GP runs. We can see that there is significant variation in the best fitness and complexity of the individuals developed by the different GP runs. However, in general the improvement in best fitness appears to level out towards the over time in all runs, confirming that the selected fixed time compute budget of 12 hours wall time is acceptable. A range of different run statistics for all 25 SSOGP runs are detailed in Appendix G.

The final symbolic regression model produced by SSOGP included 7 predictors: age (AGEn), recorded diagnosis of 'other' CVD (CVD.other.n), lipid-lowering therapy (LLTn), ethnicity (ETHNICf), SBP(SBPf), LDL (LDLf), and triglycerides (TRIGf), in addition to the discrete time indicator (t_j), which is present in all the genetic programming models to represent the jth time interval. The final prediction model generated by genetic programming is presented in figure 7.2, which is a binary parse tree representing Equation 7.1.

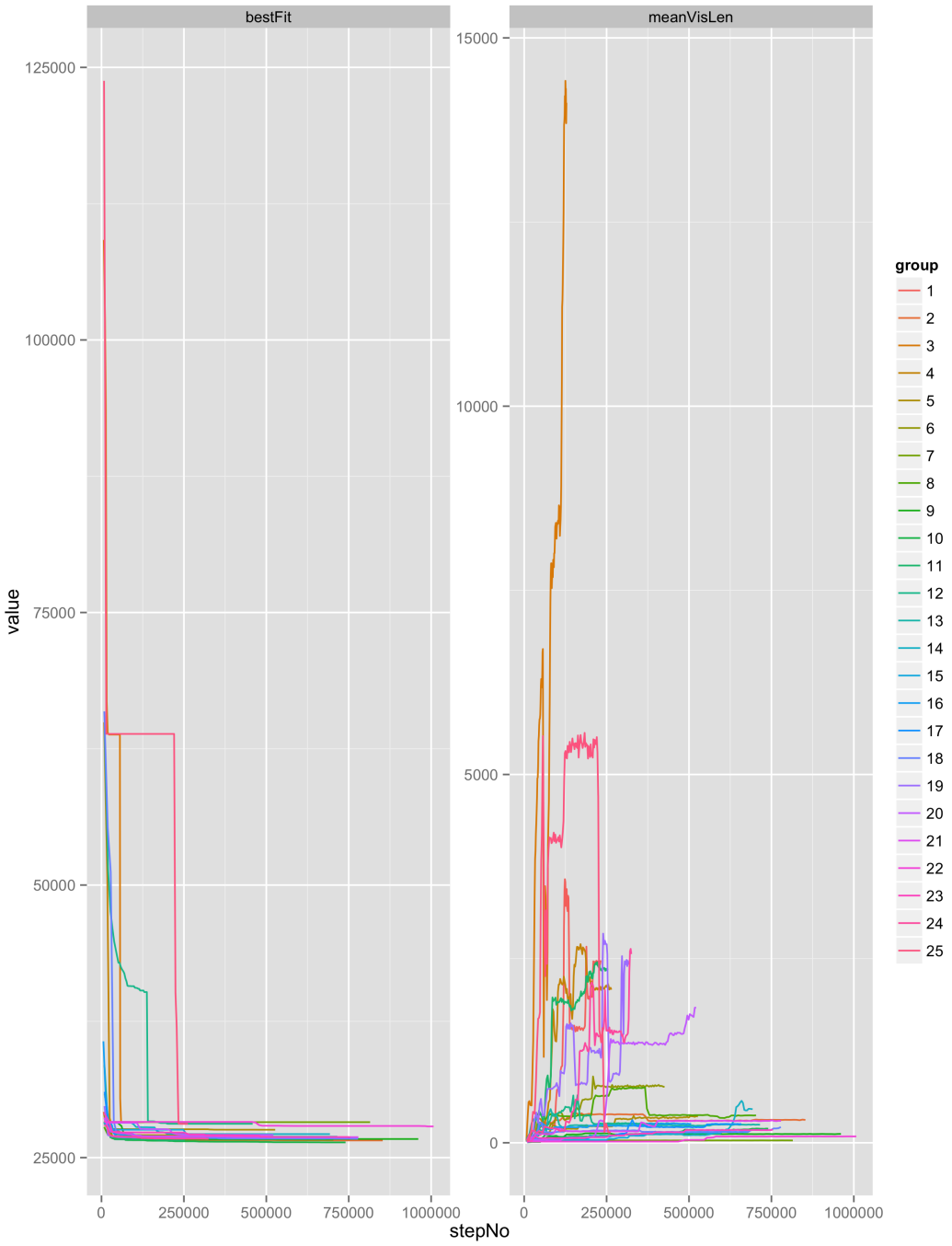


Figure 7.1: Selected runs statisitcs for the 25 SSOGP runs in the CPRD experiments.

$$\hat{\lambda}(t_j, X) = Prob(T = t_j | T \geq t_j, X) = \frac{1}{1 + e^{-X\hat{\beta}}}, where$$

$$X\hat{\beta} = \cos(\sin(\sin(ETHNICfMssn.n * -2.599 + \sin(LDLfOptm.n *$$
$$(LDLfOptm.n * -4.092 + \tan(\cos(CVD.other.n + \sin($$
$$ETHNICfMssn.n * -2.599 + \sin(LDLfOptm.n * (LDLfOptm.n *$$
$$- 4.092 + \tan(\cos(CVD.other.n + \sin(\sin(ETHNICfMssn.n *$$

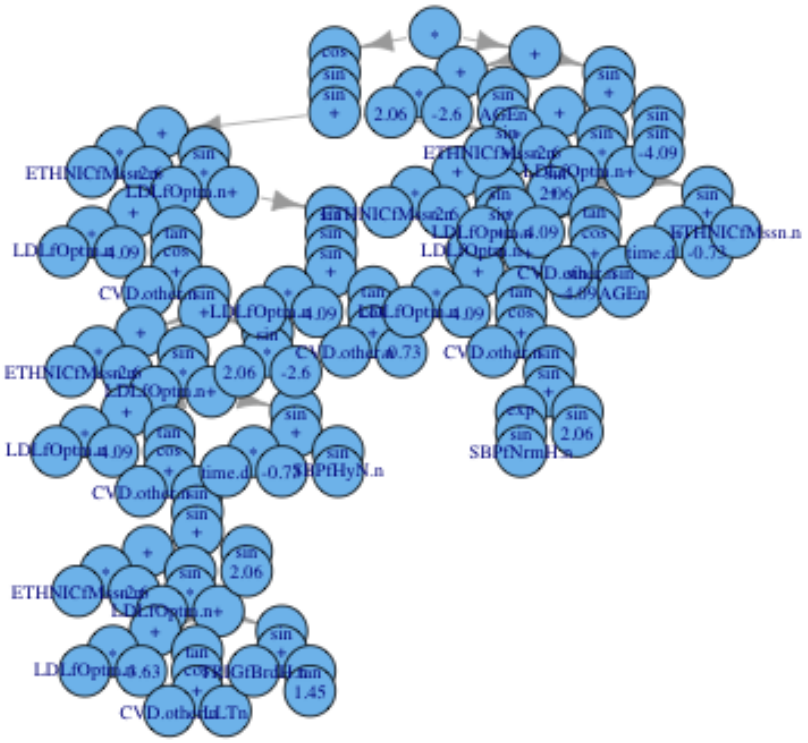


Figure 7.2: The final model developed by genetic programming, presented as a binary tree.

The other 24 prediction model generated by genetic programming are presented in Appendix H.

The genetic programming approach was applied 25 times, each time trained and tested on a different stratified re-sample of the derivation data set. This leads to a pool 25 different 'best of run' models, each of which may have selected different subset of predictors as inputs and as such may have differing levels of performance. In this pool of genetic programming models, the mean number of predictors used was 5 (Inter-

Quartile Range (IQR): 4—7). The backwards step-wise selection procedure used in the Cox modelling was also repeated 25 times, using bootstrap re-sampling to better understand the frequencies at which different subsets of predictors were selected. In the pool of 25 backwards selected Cox models the mean number of predictors used was 20 (IQR: 19—21). There was a reasonable association between the estimated effect of a predictor according in the reduced backwards step-wise model and the frequency of the selection when the step-wise selection was repeated in the bootstrap procedure (table 7.5).

Generally the features selected by repeating the GP were far more variable than the features selected by the Cox regression stepwise selection procedure (table 7.5). This is to be be expected as GP is a stochastic system, where as the stepwise procedure is deterministic, only giving variable results because we are repeating the procedure on different bootstrap resamples of the derivation data set. Despite this, the predictors that were estimated to have the largest effect in the final stepwise selected Cox model which were also selected with the high frequency in bootstrapped stepwise selection—age, recorded diagnosis of 'other' CVD, and ethnicity—were selected (relatively) frequently when GP was repeated. Interestingly, stepwise selection also often selected ACE/ARB therapy, atrial fibrillation, APT therapy, Beta-blocker therapy, BMI, Charlson index, No. GP attendances year prior, HDL, diagnosis of renal disease, SBP, SES, gender, smoking status, TC, and triglycerides as predictors, however, these featured in a low proportion of the GP models. However, SBP and triglycerides did feature in the best performing 'final' GP model. Conversely, the final GP model featured lipid-lowering therapy and LDL as predictors, predictors that was estimated to have small effect s and were selected in low proportion of Cox models by the stepwise selection procedure.

7.3.3 Model Validation

labelsubsec:cprd_validation Using the validation data set, the average performance of the 25 'best of run' prediction models automatically generated by genetic programming

Table 7.5: Number (proportion) of times predictors were selected duting the 25 repititions of Cox regression backwards step-wise selection procedure and genetic programming.

Predictor		Cox Regression	Genetic Programming
ACE/ARB Tx	ACEARB	25 (1.00)	1 (0.04)
Atrial Fibrillation	AF	25 (1.00)	4 (0.16)
Age (at baseline)	AGE	25 (1.00)	20 (0.80)
Anti-Platelet Tx	APT	25 (1.00)	9 (0.36)
Beta-blocker Tx	BETAB	23 (0.92)	0 (0.00)
Body mass index	BMIf	22 (0.88)	0 (0.00)
Charlson index	CHARLS	23 (0.92)	2 (0.08)
No. GP attendances year prior	CONT	25 (1.00)	1 (0.04)
Recorded diagnosis of Other CVD	CVD.other	25 (1.00)	19 (0.76)
Ethnicity	ETHNICf	25 (1.00)	20 (0.80)
High-density lipoprotein cholesterol	HDLf	20 (0.80)	4 (0.16)
Treated hypertension	HYPER	6 (0.24)	1 (0.04)
Low-density lipoprotein cholesterol	LDLf	14 (0.56)	7 (0.28)
Lipid lowering Tx	LLT	13 (0.52)	3 (0.12)
Rheumatoid arthritis	RA	16 (0.64)	0 (0.00)
Renal disease	RENAL	24 (0.96)	2 (0.08)
Systolic blood pressure	SBPf	24 (0.96)	2 (0.08)
Index of Multiple Deprivation	SES_5	25 (1.00)	1 (0.04)
Gender	SEX	25 (1.00)	3 (0.12)
Smoking status	SMOKf	25 (1.00)	9 (0.36)
Duration of T2DM	T2DM.dur	14 (0.56)	0 (0.00)
Total cholesterol	TCf	25 (1.00)	5 (0.20)
Triglycerides	TRIGf	25 (1.00)	7 (0.28)

was compared with the calibrated final Cox model. Graphical comparisons of the $S(t)$ values produced by each model with those obtained by the KM method in the validation set are shown in figure 7.3. As can be seen from this figure, both the Cox and genetic programming models produced similar values that had good agreement with the KM estimates in the earlier years. However, this agreement deteriorated in the latter years, where the KM estimates have high variability, as indicates by the large error bars. This high variation my be explained by the fact that with a median follow-up time of 5.2 years, there are far fewer events and number of subjects in the latter time periods. Whilst agreement deteriorated in the latter time-points, both models had generally acceptable overall agreement.

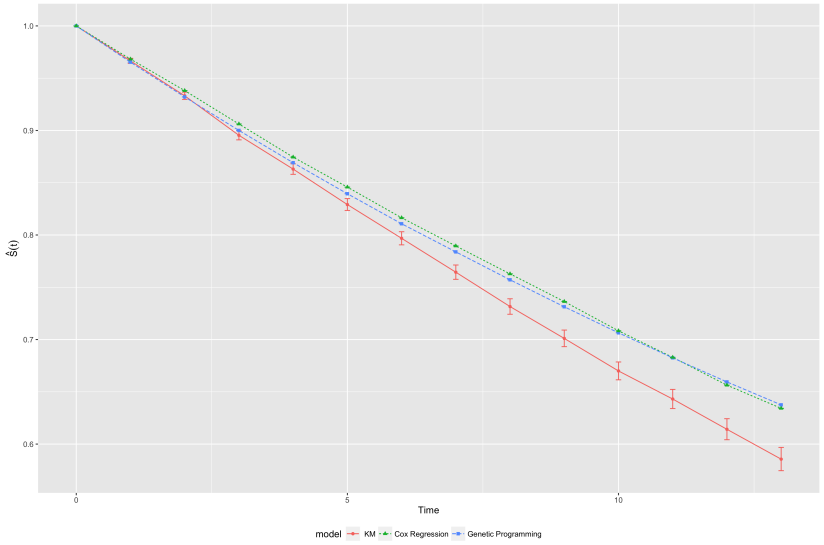


Figure 7.3: Average survival curves for the Cox regression and genetic programming models. The error bars represent ± 2 standard errors of the KM estimates.

The discriminative performance in the validation set, according to the C-statistic, of the models at different time points is shown in table 7.6 and figure 7.4. From the C-statistic estimates we can see that a satisfactory performance of >0.6 was reached in both models, at all time points. There was generally comparable discriminative performance of both models, with the Cox model showing marginally better discrimination at all time points.

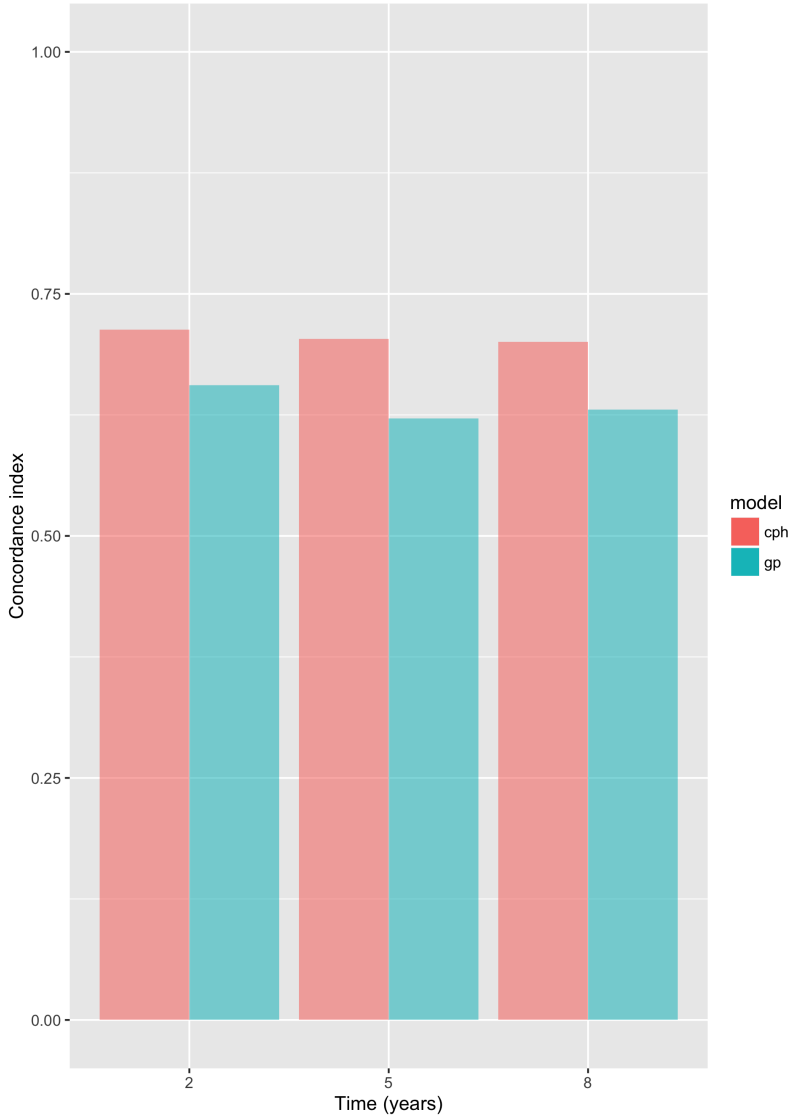


Figure 7.4: C-statistic estimates by model for t=1, 3 and 5 years

The calibration plots evaluated by grouping subjects according to quantiles of predicted risk ($1 - S(t)$) at t = 2, 5, and 8 years are shown in figure 7.5. From the graphical inspection of the calibration plots we can see that there was no tendency to systematically over- or under-predict at any of the time points in either the Cox or genetic programming models. Again from visual inspection, the genetic programming model appeared to be marginally less calibrated than the Cox model, however, both models demonstrated comparable performance with significant lack of calibration at any time

Table 7.6: C-statistic estimates by model at t=2, 5, and 8 years

Time (years)	Cox PH Regression	Genetic Programming
2	0.713	0.656
5	0.703	0.621
8	0.701	0.631

point.

The corresponding χ^2 statistics and p-values are shown in table 7.7 for completeness only. As with any statistical test, the power increases with sample size; this can be undesirable for goodness of fit tests because in very large data sets, small departures from the proposed model will be considered significant. Because of the very large sample sizes studied here, a statistically significant Hosmer-Lemeshow statistic (as detailed in table 7.7) is not considered informative with respect to calibration [81, 235].

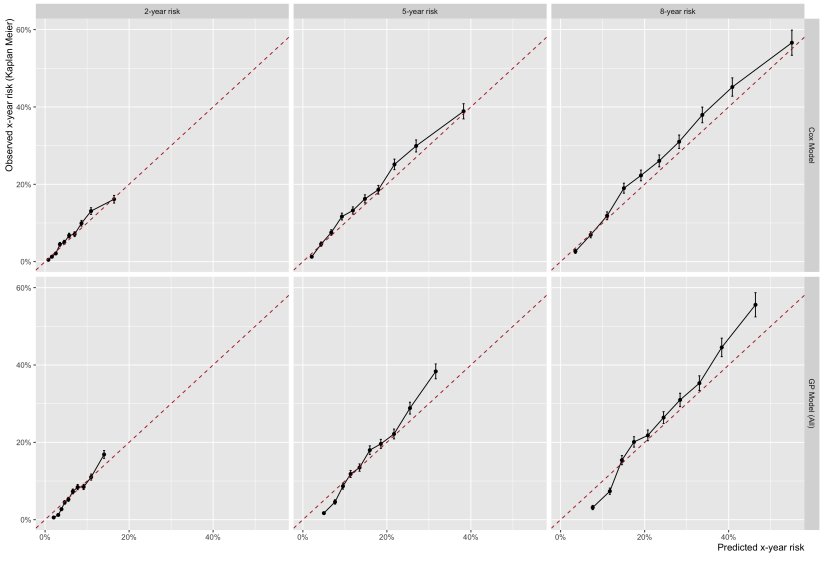


Figure 7.5: Calibration plots for the Cox regression and genetic programming models, at t=2, 5, and 8 years..

Table 7.7: χ^2 statistic for the comparison between observed versus expected (according to the model) number of events in groups of patients defined according to the predicted $1 - S(t)$ at t=2, 5, and 8 years.

Time (years)	Cox Regression		Genetic Programming	
t	χ^2	p-value	χ^2	p-value
2	1589	< 0.001	1575	< 0.001
5	4612	< 0.001	4146	< 0.001
8	8236	< 0.001	6937	< 0.001

7.4 Discussion

This study showed that Cox regression and GP produced similar results when evaluated in a common validation data set. After re-calibration the discriminative ability of the Cox model in the validation set was slightly larger than that of the GP model at all time points. Despite slight relative differences, both models demonstrated an acceptable level of discriminative ability (C-index >0.6) at all times points. The GP model had marginally poorer calibration when visually compared with the Cox model. However, both models demonstrated no significant lack of calibration at any time point.

Despite generally comparable performance, albeit in favour of the Cox model, the predictors selected for representing their relationship with the outcome were quite different. The final reduced Cox model used 20 predictors, in contrast to 7 predictors used in the GP model. The GP model used significantly fewer predictors, further confirmed by repeating the the GP and the stepwise selection procedure used in the Cox modelling, resulting in mean numbers of predictors of 5 (IQR: 4—7) and 20 (IQR: 19—21), respectively.

Predictors that were estimated to have larger contributions to the final Cox model and that frequently selected during stepwise selection—age, recorded diagnosis of 'other'

CVD, and ethnicity—were selected in the final GP model and with (relatively) high frequency when the GP was repeated. However, others predictors that had large to moderate contributions to the final Cox model—ACE/ARB therapy, atrial fibrillation, APT therapy, Beta-blocker therapy,BMI, Charlson index, No. GP attendances year prior, HDL, diagnosis of renal disease, SBP, SES, gender, smoking status, TC, and triglycerides —were selected infrequently when GP was repeated. However, SBP and triglycerides did feature in the best performing 'final' GP model. Conversely, the final GP model also featured lipid-lowering therapy and LDL as predictors, predictors that was estimated to have small effect s and were selected in low proportion of Cox models by the stepwise selection procedure. Whilst these results confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling, symbolic regression model did not estimate such a large number predictors to be strongly associated with the outcome, associated strongly enough for inclusion in the model at least, whilst achieving comparable performance.

As with the experiments in the previous chapter, these results suggest that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors. To test the first part of this hypothesis—that GP can represent the potential non-linear relationships that exist between predictors—the shape of the predictor effects were plotted to evaluate whether or not the effects were non-linear in nature. Using the 'final' GP model, the effects of each predictor's values were plotted against log hazard, whilst the other values were held at their reference values. Reference values were the modal class for binary variables and the mean of continuous variables. Figure I.1 of appendix I illustrates that the the 'final' model developed by GP is modelling non-linear effects for the continuous predictor age (which was the only continuous predictor in the model, all the others were binary).

To test the second part of this hypothesis—that GP can better represent these relationships between predictors using fewer variables—we repeated the GP runs with exactly the same experimental set-up, but restricting the inputs to predictors that were selected

with a relatively high frequency (>0.5) in the original GP run of the primary experiment. The covariates included age, recorded diagnosis of other CVD, and ethnicity. This produced very similar results (detailed in appendix J) both in terms of calibration and discrimination. Both these findings support the hypothesis that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

Whilst considerable effort was made to relax the linearity of the Cox regression, through transformation of predictors, the nature of the approach relies on linear combinations of predictors. The fact that GP required fewer predictors to achieve similar performance may have an advantage in practical application of the developed clinical prediction model. The acquisition of information that forms the inputs to such a model can be prohibitively onerous in routine clinical practice. Therefore a prediction model that requires fewer inputs, especially if the information relating to these inputs is in practice recorded easily and to a good quality, would considerably increase adoption and utility.

This work has limitations introduced by its use of data from CPRD to refine a cohort of patients, a cohort from a primary prevention setting, designed to predict the risk of primary cardiovascular events in patients withT2DM who have presented with clinical CVD. Through its use of the CPRD data this work has demonstrated the utility of GP in a primary prevention setting, however, there are limitations in the generalisability of these findings to the other clinical settings of cardiovascular risk prediction. However, its important to note primary prevention in asymptotic cardiovascular patients is arguably the most common clinical setting for the application of cardiovascular risk prediction models in routine practice.

7.5 Conclusion

Using data from CPRD we demonstrated that a symbolic regression model developed by SSOGP has predictive ability comparable to that of Cox regression for the predic-

tion of future cardiovascular events in patients with T2DM in a primary prevention setting, i.e in T2DM patients with asymptomatic cardiovascular disease. These experiments compared an untuned SSOGP symbolic regression model that was developed in an automated fashion using basic parameter values recommended from the GP literature, with a highly tuned Cox regression model that was developed in a very involved manner that required a certain amount of clinical and statistical expertise. Whilst the highly tuned Cox regression model was better calibrated to validation data and the untuned genetic programming model had better discriminative ability, the performance of the automatically generated prediction model was generally comparable. These findings confirm those of the previous experiments and demonstrate the utility of GP as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes, where the primary goal is accurate prediction. These findings also confirm the prognostic significance of age, recorded diagnosis of 'other' CVD, and ethnicity, and to a lesser extent SBP and triglycerides, for cardiovascular risk in T2DM patients with asymptomatic CVD.

In the next chapter the hypotheses and goals of this thesis are revisited and contributions discussed in light of the results of this work. These results are summarised, discussed, and put into context. Limitations of this work are critically assessed and opportunities for further research identified.

Chapter 8

Discussion & Conclusions

The previous chapters have described the wider context clinical prediction modelling and the UK health system, defining the challenges of predicting risk in the presence of censored data, and providing motivation for the application of GP for cardiovascular risk prediction. Then we surveyed and critically assessed the existing research related to this thesis. Next we gave an overview of the essential common themes in the diverse field of GP and discussed the specific methodological elements that formed the developed GP approach for censored longitudinal data, which was implemented and assessed in the subsequent experiment chapters. Then we performed our first set of experiments that independently and externally validated the performance of the *de facto* cardiovascular risk prediction model for patients with T2DM, the UKPDS-RE, using data from CPRD. The results of these experiments showed poor performance, suggesting that the UKPDS-RE is not suitable for predicting cardiovascular risk in UK subjects with T2DM and that there is a need for revised risk models in T2DM. Next we discussed our second set of experiments that demonstrated the utility of the developed GP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events in patients with symptomatic cardiovascular disease using censored survival data from the SMART study. Finally, we discussed our third and final set of experiments with a very similar experimental set-up those in the previous chapter, that demonstrated the utility of the developed GP approach for the automatic development of clinical prediction models for risk prediction of future cardiovascular events, but used a much larger observational cohort of patients from CPRD

in a primary prevention clinical setting, where patients have asymptomatic CVD.

In this chapter the hypotheses and goals of this thesis are revisited and contributions discussed in light of the results of this work. These results are summarised, discussed, and put into context. Limitations of this work are critically assessed and opportunities for further research identified.

8.1 Contributions of this Work

This thesis makes six main contributions:

1. *The de facto cardiovascular risk prediction models for T2DM may be unsuitable* Using data from CPRD this work has preformed the largest, independent, external validation of the *de facto* cardiovascular risk model for people with T2DM, the UKPDS-RE, in a diverse and contemporary setting. This work showed poor performance, suggesting that the UKPDS-RE is not suitable for predicting cardiovascular risk in UK subjects with T2DM. Considering the widespread application of these prediction models, this work suggests a need for revised risk equations in T2DM.

2. *Developement of a GP approach for survial analysis of censored data* GP is a general methodology, the specific implementation of which requires development of several different specific elements such as problem representation, fitness, selection and genetic variation. This work has developed a tree-based untyped SSOGP approach for the automated development of clinical prediction models in the presence of censored longitudinal data. Specific GP elements were developed and implemented, such as fitness functions and search heuristics, to handle the problem-specific complexities of censored data and facilitate survival analysis.

3. *Generational GP approaches are too computionally expesive for large observational cohorts* This work attempted to implement and evaluate the utility of two broad classes of GP, steady-state GP common in modern GP systems and the more traditional generational GP approach. Despite considerable effort, when the developed generational approaches were applied to the large observational datasets of censored longitudinal data identified for this work, they failed as a result of requiring more memory that was available in the computing resources allocated for this work. This serves to demonstrate the utility of the relatively computationally efficient steady-state GP approach for analysing large observational cohorts of patients.

4. *GP has utility for the automatic development of clinical prediction models in censored data* Using data from the SMART study and from CPRD we have demonstrated that symbolic regression models generated by the developed SSOGP approach had predictive ability comparable to that of the *de facto* statistical method—Cox regression—for the prediction of future cardiovascular events in patients with symptomatic and asymptomatic CVD. These experiments compared untuned SSOGP symbolic regression models that were developed in an automated fashion using only basic parameters settings recommended from the GP literature, with highly tuned Cox regression models that were developed in a very involved manner that required a certain amount of clinical and statistical expertise. Whilst the highly tuned Cox regression models performed slightly better in their validation datasets, the performance of the automatically generated symbolic regression models were generally comparable, and on average consisting of considerably fewer predictors. Using symptomatic and asymptomatic CVD as case studies—for secondary and primary prevention clinical settings, respectively—these findings demonstrate the utility of GP as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes in the presence of censored longitudinal data.

5. *Confirmation of the prognostic significance of certain risk factors in symtomatic CVD* This work has applied GP to examine the prognostic significance of different risk factors together with their non-linear combinations in predicting cardiovascular outcomes in patients with symptomatic and asymptomatic CVD. Whilst the application of GP did not provide more accurate representations of factors that predict the risk of both symptomatic and asymptomatic CVD when compared with existing methods, GP did offer comparable performance. Despite generally comparable performance, albeit in slight favour of the Cox model, the predictors selected for representing their relationships with the outcome were quite different and, on average, the models developed using GP used considerably fewer predictors. The results of the GP confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling; age, previous atherosclerosis, and albumin for secondary prevention; age, recorded diagnosis of 'other' CVD, and ethnicity for primary prevention in patients with T2DM. When considered as a whole, GP did not produce a better performing clinical prediction model, rather it utilised fewer predictors, most of which were the predictors that the Cox regression estimated be most strongly associated with the outcome, whilst achieving comparable performance. This suggests that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

6. *In practice GP is robust* By implementing SSOGP without model tuning, using only basic parameters values recommended from general GP literature, observing that it has performance comparable to the *de facto* statistical method, we have confirmed the observations of other authors, that in practice GP is robust and likely to work well over a wide range of parameter values.

As stated in the introduction, the main hypothesis of this research is that the application of GP can provide more accurate representation of factors that predict the risk of cardiovascular disease when compared with existing methods. This work repres-

ents a successful first attempt at evaluating this hypothesis. The results of this work may not be able to confirm conclusively whether GP offers a more accurate representation of factors that predict the risk of cardiovascular disease when compared with Cox regression. However, they can confirm that GP offers comparable accuracy, whilst developing clinical prediction models in an automated fashion that require fewer predictors.

Specifically, in-line with main goals of this work described in the section 1.2, this work has provided evidence that the *de facto* cardiovascular risk prediction models for T2DM may be unsuitable - motivating the need for improved clinical risk prediction methods and models for survival outcomes in contemporary populations. It has also demonstrated the utility of GP for the automatic development of clinical prediction models and examined the prognostic significance of different risk factors together with their non-linear combinations, using two different CVD case studies. This work has successfully achieved these three main goals in contributions one, four, and five, respectively.

8.2 Discussion

GP is a general methodology, the specific implementation of which requires development of several different specific elements such as problem representation, fitness, selection and genetic variation. In chapter 4 we developed a tree-based untyped SSOGP approach for the automated development of clinical prediction models in the presence of censored longitudinal data. Specific GP elements were developed and implemented, such as fitness functions and search heuristics, to handle the problem-specific complexities of censored data and facilitate survival analysis. In fact, single and multi-objective generation GP approaches (GSOGP and GMOGP) were also developed. However, the utility of these approaches could not be assessed because when implemented using our experimental set-up, these generational approaches proved too memory intensive for

the computing resources available (42Gb memory).

Problem representation was addressed by taking advantage of the fact that the hazard function corresponds to a conditional probability in the discrete time domain, casting the original survival analysis problem into a classification problem that required the estimation of a conditional probability. However, to address the problem of censoring the data needed to be pre-processed into the *counting process format*, with multiple rows per subject, one for each observed discrete-time interval. A suitable fitness measure was developed for symbolic regression in the presence of censored survival data based on ML estimation to calculate the distance between the natural log of the predicted probability of the event to the actual observed outcome. The developed fitness function expresses the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown population parameters. In the absence of suitable existing GP system, the developed GP approach was implemented from scratch using the R statistical programming language. However, it was only implemented from scratch in the general sense, i.e. we did not use an existing GP system. Existing GP- and EA-specific R packages provided a significant amount of problem-agnostic functionality required to implement our GP system. However, in implementing a GP system for survival analysis there was significant development required, mainly with respect to the implementation a suitable fitness function for censored data, modification of search operators, and heuristics to work with the specific problem representation and associated counting process data format.

In chapter 5 we evaluated the performance of the *de facto* cardiovascular risk for people T2DM, the UKPDS-RE, for predicting the 10-year risk of CVD in a cohort of UK patients from CPRD newly diagnosed with T2DM. At the time of writing, this work is the largest, independent, external validation of the UKPDS-RE, in a diverse contemporary setting. The four UKPDS risk equations constituting the UKPDS-RE showed a reasonable ability to identify high-risk patients (discrimination) but were generally poor at quantifying the absolute risk (calibration). The UKPDS-RE CHD risk equations

consistently overestimated absolute risk, whereas the UKPDS-RE stroke equations performed relatively well. However, when considered as a whole, the UKPDS-RE was unsuitable for predicting CVD risk in UK subjects with newly diagnosed T2DM. These findings suggest that the use of UKPDS-RE in clinical practice will lead to over-estimation of CVD risk in patients with newly diagnosed T2DM. This in turn is likely to lead to selection of preventative treatments, for which, for some patients, the balance of risks may outweigh the benefits. Considering the widespread application of these prediction models in clinical practice, drug reimbursement, and public health decision-making, these results suggest that there is a need for revised risk equations in T2DM.

In chapters 6 and 7 we compared the performance of the *de facto* statistical method for survival analysis, Cox regression, with the developed SSOGP approach in the development of clinical prediction models for the prediction of future cardiovascular events in patients with symptomatic and asymptomatic cardiovascular disease, using data from the SMART study and CPRD, respectively. Both these experiments showed that Cox regression and the developed SSOGP approach produced similar results when evaluated in common validation datasets. Despite slight relative differences, both approaches demonstrated an acceptable level of discriminative and calibration at a range of times points.

Whilst the application of GP did not provide more accurate representations of factors that predict the risk of both symptomatic and asymptomatic CVD when compared with existing methods, GP did offer comparable performance. Despite generally comparable performance, albeit in slight favour of the Cox model, the predictors selected for representing their relationships with the outcome were quite different and, on average, the models developed using GP used considerably fewer predictors. The results of the GP confirm the prognostic significance of a small number of the most highly associated predictors in the Cox modelling; age, previous atherosclerosis, and albumin for secondary prevention; age, recorded diagnosis of 'other' CVD, and ethnicity for primary

prevention in patients with T2DM. When considered as a whole, GP did not produce a better performing clinical prediction model, rather it utilised fewer predictors, most of which were the predictors that the Cox regression estimated be most strongly associated with the outcome, whilst achieving comparable performance. This suggests that GP may better represent the potentially non-linear relationship of (a smaller subset of) the strongest predictors.

Whilst considerable effort was made to relax the linearity of the Cox regression, through transformation of predictors, the nature of the approach relies on linear combinations of predictors. The fact that symbolic regression required fewer predictors to achieve similar performance may have an advantage in practical application of the developed clinical prediction model. The acquisition of information that forms the inputs to such a model can be prohibitively onerous in routine clinical practice. Therefore a prediction model that requires fewer inputs, especially if the information relating to these inputs is in practice recorded easily and to a good quality, would considerably increase adoption and utility.

Unlike other machine learning algorithms, GP in not a 'black box' method and provides a mathematical formula as its output. However, the model structure in the GP model is typically more complex than that of the Cox regression model. This hinders the interpretation of the (relative) effects of predictors on the outcome and if the primary objective of the modelling is to understand these effects, such as in aetiologic research, then Cox regression and other related approaches still remain the first choice. However, if the primary goal of the research is accurate risk prediction, then GP has some utility when compared with its regression counterpart. GP offers advantages that include its ability to learn complex non-linear relationships that may exist in the data, it is not confined by the statistical assumptions that underpin Cox regression (such as proportional hazards), its inherent feature selection, and that GP models are developed in automated fashion.

An advantage of Cox regression was its ability to be calibrated using all available

data by applying a shrinkage factor - an measure of the models optimism (or over-fitting) - estimated though bootstrapping or penalised regression methods. Whereas with GP we cannot estimate a shrinkage factor in the same way and need validation sample. This suggests that in cases where the data is scarce, Cox regression may be a better approach. In contrast where there the data is large, possibly with a large number of predictors and potential interactions effects, GP would have a distinct advantage. Whilst interaction effects can be modelled using regression techniques, this can be onerous and require a degree of expertise.

The considerable statistical and clinical expertise required in the development of appropriate clinical predictions should not be understated. Problems with step-wise feature selection methods are another concern; including biased R^2 values, confidence intervals for effects and predicted values that are falsely narrow, biased regression coefficients that need shrinkage, and severe problems in the presence of collinearity. Both Cox regression and step-wise selection are widely used, and widely abused, in prognostic and aetiologic research. Whilst fitting models using these techniques is relatively straightforward and intuitive, sometimes they are applied blindly without appropriate testing of the underlying assumptions. Whilst Cox regression is a powerful tool, its correct application requires a certain amount of statistical rigour and expertise from the researcher, and cannot be used in certain data if its underpinning assumptions are violated. Another weakness of Cox model is that it does not explicitly define the underlying baseline hazard, which means that technically its predictions are only valid at the time points observed in the data and that it may not appropriate for extrapolation to non-observed time points. It should be noted, however, that other regression methods for survival analysis, such as parametric survival models, can define the baseline hazard and are appropriate for extrapolation. However parametric modelling of survival is even more involved than Cox modelling, requiring greater technical expertise, and as such features far less in published aetiologic research.

The main weakness of the GP approach is that the data need to be converted into the

counting process format, which leads to large data sets and longer executions times. So whilst methodologically GP works better on large data sets, in practice the long execution times can make its use prohibitive. However, this weakness can addressed though parallel processing. GP is a method that can be described as "naturally parallelizable", and as such can adapted to execute in parallel across multiple machines or processors.

Finally, the GP model has a number of parameters that need to be specified *a priori*. These parameters include the size of the population, the building block of models such as mathematical operators, how many runs to perform, the rates at which to apply genetic variation such as crossover and mutation, and parameters such as maximum tree depth that control the complexity, and thus potential of over fitting, of final GP model. Often the choice of these parameters in based on trail and error, model tuning, or from the literature. Model tuning refers to repeating the same experiment many times whilst simultaneously varying multiple parameters and quantifying relationship between them and the quality of resultant models to understand which parameters are important.

However, there are little or no literature on the relative importance of specific parameters of survival analysis. Model tuning was outside of the scope of this work, but further research is warranted into characterising the association of GP parameters and performance in a survival analysis setting. In the absence of modelling tuning and suitable literature, we used arbitrarily selected model parameters that were in an order of magnitude with widely accepted starting parameters of GP applied in other settings.

8.3 Critical Assessment

There are typically a number of limitations of when using real-world observational data, and the experiments in this thesis are no exception. Data have been collated from routine clinical practice, thus there are missing and erroneous data, coding imperfections, lack of standardisation of biochemical measures (such as lipid profiles), vari-

ations between biochemical test centres and measurements are taken with varying periodicity. Measurement error in identifying the CVD outcomes will have been present in the analysis, but this work has endeavoured to select appropriate medical codes for the cardiovascular endpoints involved, consulting clinical expertise whenever possible. Certain covariates of interest such as smoking status, BMI, lipids, family history of CVD, etc, may not be recorded consistently. There are also limitations with ethnicity data where even within the non-missing data there are a large proportion of ethnicities recorded as 'unknown'. Removal or exclusion of patients with missing data may introduce bias into the study; this was addressed by using multiple imputation techniques to impute missing values or modelling as categorical variables with a missing level/category, where appropriate. There are also limitations in specifying covariates or predictors *a priori* as there is potential to miss important factors and relationships that exist with variables not considered. There are limitations on the split-sample validation approach where predictive accuracy estimates, although unbiased, can be imprecise. There are also limitations inherent in the classic statistical modelling techniques used in this work, each of which have their own set of assumptions, such as non-informative censoring, linearity, additively, proportionality, etc., that need to be satisfied in order to order to take a given approach. Violation of such statistical assumptions may have precluded the use of certain techniques and/or consideration of all covariates.

A key strength of the experiments in this thesis have been the size of the datasets utilised and study design adopted. We have used real-world observation medical datasets with a retrospective cohort study design to minimise the selection bias by using real patient data, observed in routine clinical practice, that has not been subject to the selection biases inherent in other types of data and study designs, such those associated with RCT data. We have used relatively large datasets which are important to give adequate statistical power to detect the associations with the outcome, something even more important in the context of CVD, where generally events are quite rare and thus large sample sizes are required. However, the strength of the these large datasets has introduced key challenges from the resultant increase in computational expense required

to analyses them. Potentially, smaller datasets could have been used, or subsets of the selected datasets, however this was not done because of the resultant loss of statistical power to detect the associations between the predictors and the outcome, which in the context of this thesis are rare cardiovascular events.

The choice of implementation language, the R statistical programming language, did not help with computational issues, as a key disadvantage of R is its performance for computationally expensive tasks. However, it was felt that the disadvantages of the Rs computational inefficiency were outweighed by the advantages of how it internally represents expressions as trees, supporting direct manipulation of expression trees through the same syntax for manipulating nested lists, making implementation of GP operators in R simple, succinct, and easy for proficient R users to understand.

The computational issues from the size of the data were further exasperated by requirement, by the fitness function and other elements of the approach developed for this work, to have the data in the counting process or long format. This meant that rather than requiring one row per subject, multiple rows per subject—one for each time segment for which the subject was observed—were required. This requirement significantly increases the size of the data and the number of fitness cases that needed to be evaluated, in turn, significantly increasing the computational expense of the analysis.

Another limitation is the use of the death penalty for constraining the search space. The literature suggests that this approach is generally not the best and that, in the first instance, measures that penalise individuals based on some degree of invalidity or infeasibility are superior. The rationale was that the death penalty is computationally cheap when compared to other methods and that due to the nature of the experiments in this that computational expense had become an important consideration.

At the outset it was the intention to evaluate steady-state and generational, as well as single- and multi-objective approaches to symbolic regression for clinical prediction modelling of censored survival data. Whilst Steady-state Multi-Objective Genetic

Programming (SMOGP) was not supported in the R packages selected for this work, SSOGP, GSOGP and, GMOGP were. However, during the experiments the generational approaches (GSOGP and GMOGP), were unsuccessful due to the fact that they exceeded the available memory when applied to the large datasets selected for the experiments. As a result, only models developed using SSOGP we able to be evaluated in this work. The fact that only single-objective GP was evaluated in a key limitation of this thesis.

Using only a simple single-objective steady-state search heuristic (i.e. SSOGP) there wasn't any mechanism to preserve genetic diversity in the population. A multi-objective search heuristic would have helped with controlling the size and complexity of the solutions developed and thus controlling bloat, however this has been controlled for to a lesser degree, buy constraining the maximum depth of new individuals or subtrees developed from the initialisation and variation operators. A SMOGP search heuristic was not developed from scratch, for which a significant amount of work would have been required to integrate it with the existing R packages used for the problem agnostic elements, as it would have been beyond the scope of this work.

Another limitation of examining only steady-state GP is the noise that is introduced thought the random selection of individuals in tournament selection. Meaning that individuals with average fitness can have some chance of being selected and their offspring featuring in future generations. However, steady-state GP is far more computationally efficient, which was major consideration when analysing large observation datasets. This is reflected in the fact that whilst this work attempted to evaluate single and multi-objective generational approaches, they were too computationally expensive to be used the large datasets identified for this thesis. The fact that the utility of generational GP approaches for clinical prediction modelling in censored survival data was not evaluated is another limitation of this work.

In this work we have chosen to consider the most common type of GP, untyped tree-based GP. A limitation of this work is that we have not characterised the utility of

other types of GP, such as linear or graph-based GP. Evaluating other types of GP was considered outside the scope of this work as we wanted to access the utility of the most accessible and most researched type of GP with basic parameters settings, rather than evaluate more specialised forms of GP where the literature may be more sparse. For this thesis we enforced closure and defined the valid search space through fitness penalisation, which is a computationally less efficient approach, as many invalid solutions would have been created and not considered by giving them the lowest possible fitness. A more efficient approach may have been to implement a typed GP system to explicitly define the valid search space and avoid the evaluation of invalid solutions. However, this does not mean that the not excluding invalid search spaces would have lead to better solutions, as these invalid regions can form an important intermediate steps towards a (near-)optimal solutions, just that in this case a more computationally efficient GP implementation would have reduced the required resources and possibly enabled a wider range of experiments to be performed.

Another limitation of this work is that it did not characterise the association of GP parameters and performance in a survival analysis setting. This is because model tuning was outside of the scope of this work. In the absence of modelling tuning and suitable literature, we used arbitrarily selected model parameters that were in an order of magnitude with widely accepted starting parameters of GP applied in other settings. However, the literature suggests that in practice GP is robust and likely to work well over a wide range of parameter values.

8.4 Further Work

A key weakness of the GP approach for censored data presented in this work is that the data needed to be converted into the counting process format, which leads to even larger data sets, longer executions times and increased memory requirements. So whilst methodologically GP works better on large data sets, in practice the long execution

times can make its use prohibitive. Further work is required to reduce the computational expense of GP, both in terms of reducing execution times and in the reduction of hardware requirements. Speeding up GP runs and reducing memory requirements—so that analyses on large datasets could be done on standard commodity hardware in an amount of time in the same order of magnitude as classical regression methods—would make GP far more accessible to a broader audience, facilitate, and expedite research into extensions and applications of GP in the big data era.

Reduction in the computational expense of GP may enable another area of potential further work, the characterisation of the utility of generation GP approaches for clinical prediction modelling in the censored data. In this work using large medical datasets, the memory requirements of the generational search heuristics exceeded the available computing resources, even using compute nodes with 42Gb of memory. Although beyond the scope of this work, these issues of execution time and hardware requirements can addressed though parallel processing. GP is a method that can be described as 'naturally parallelisable', and as such can adapted to execute in parallel across multiple machines or processors. The landscape of big data analytics has changed dramatically in the last few years, a trend that looks set to continue for some years to come, with exciting new technologies such as *Hadoop* and *Spark* facilitating highly distributed computing on commodity hardware.

Also, an approach for performing GP on censored survival data that does not require the data to be pre-processed into the long format would be advantageous in terms of reducing the computation expense. However, it must be noted that although not explored in this work, the long format enables the analysis of time-varying covariates which can be highly advantageous when answering certain prognostic research questions, another potential avenue for further research. The development of other fitness measures and constraint handling approaches specific to survival analysis would be an important area for further research. It would be very interesting to compare their performance and understand how these effect the search and the resultant solutions.

Another area of further work would be assess the impact of using a multi-objective steady-state GP on the performance of clinical prediction models developed on censored survival data. It would be interesting to understand how controlling genetic diversity and bloat though additional jointly-optimised objectives would effect the performance of the survival models developed by GP, and whether they could outperform the *de facto* statistical methods for survival analysis.

Whilst this work supports the observations of other authors, that GP is robust and works well on a range parameter settings, further work is required to characterise relationship between GP run parameters and the performance of the resulting survival models. Model tuning was outside of the scope of this work, however, characterising the relative effects of GP parameter settings on the performance of the developed clinical prediction models may have a number of benefits. By understanding which parameters do not drive performance we could potentially reduce the computational expense by constraining some aspect of the experimental design or search space, such as reducing the number of fitness evaluations or reducing the number and average size of individuals. By understanding the near-optimal GP parameter settings for the development of clinical prediction models in censored survival data, GP may be able to provide more accurate representation of factors that predict the risk of cardiovascular disease and demonstrate improved performance when compared with existing methods.

Finally, further work is required to assess the utility of the developed GP approach for automated development of new clinical prediction models in other clinical and environmental settings, preferably comparing their performance against more established risk prediction models currently used in routine clinical practice (for example QRISK 2 [125] and SCORE [46] for primary prevention of CVD, the GRACE score [99, 63] in acute coronary syndromes and, euroSCORE [215] in Cardiac surgery). Further work will also be required to identify the most appropriate clinical parameters required for GP risk modelling in order to optimise their predictive power in the relevant setting and to ensure that these required measures are not only practical to measure at scale in

routine clinical practice but also cost effective.

8.5 Conclusions

To our knowledge, this work is the first empirical study to assess the value of GP for clinical prediction purposes compared to the well-known and widely applied Cox PH regression technique. Using real-world data this work has demonstrated that symbolic regression models developed by SSOGP have predictive ability comparable to that of Cox regression models. The symbolic regression models were more complex and thus more difficult to validate by domain experts, however these models were developed in an automated fashion, using considerably fewer input variables, without the need for domain specific knowledge and expertise required to appropriately perform survival analysis. GP has demonstrated strong potential as a methodology for automated development of clinical prediction models for diagnostic and prognostic purposes.

This work compared untuned SSOGP symbolic regression models that were developed in an automated fashion with highly tuned Cox regression models that were developed in a very involved manner that required a certain amount of clinical and statistical expertise. Whilst the highly tuned Cox regression models performed slightly better in validation data, the performance of the automatically generated prediction models were generally comparable. The comparable performance demonstrates the utility of GP for clinical prediction modelling and prognostic research, where the primary goal is accurate prediction. In aetiological research, where the primary goal is to examine the relative strength of association between risk factors and the outcome, then Cox regression and its variants remain as the *de facto* approach.

In hypothesis-driven research, the user formally specifies some idea—the hypothesis—*a priori* and uses data to prove (or disprove) this idea. In contrast, with a data-driven approach the user does not formally specify a hypothesis *a priori*, but uses the data to discover patterns and relationships, thus generating hypotheses. Frequentist statist-

ical approaches for clinical prediction modelling (such as Cox regression) are typically hypothesis-driven. Typically the user specifies the variables of interest and the general model form (e.g. Cox regression) *a priori* . There are ways in which these methods can be more data-driven, such as considering a broad set of variables and using stepwise feature selection procedures to select the subset of important predictors. However, as discussed previously in this chapter, there are problems with stepwise feature selection procedures and this is only partially data-driven, the general model form has already been specified *a priori* and limited to linear combinations of predictors. In contrast GP is a data-driven approach, it automatically solves problems without requiring the user to know or specify the form or structure of the solution in advance. In addition to efficiently searching for potentially complex non-linear associations between predictors and outcome, GP also searches for the optimal model form (or structure) and—intrinsic to its evolutionary process—automatically selects a subset of features.

In an era of evidence-based medicine and shared decision-making, medicine and healthcare is becoming increasingly data-driven, so too increases its overlap with the field of computer science. With increasing amounts of medical data becoming available through EPR systems and consumer-facing wearables, the opportunity for computer science to significantly reduce the burden of disease continues to grow. Computer science can be used to extract knowledge from large datasets and use it to ultimately improve both the services provided and the quality outcomes for the patient. For example, accurate risk prediction models can extracted from the data for a potentially unlimited number of disease and therapeutic areas, helping to understand in advance what course of action is more likely to have the desired outcome and how to spend resources effectively. Models can be automatically and efficiently retrained and recalibrated to reflect new tests, new interventions, new patients, and any other updates to the underlying data. This becomes more advantageous as both the volumes of medical data and efficient mechanisms for data collection continue to increase. Therefore, automatically generated and updated models are more likely to remain relevant and valid for longer, and thus be more cost-effective to develop and maintain, when compared

with hand-crafted models. Routine application of automated clinical prediction models could be used to screen entire populations and offer timely targeted interventions to those identified to be at risk, thus having potential economic and quality of life impact for society.

Appendix A

ISAC Protocol: UKPDS-RE Validation

ISAC APPLICATION FORM
OTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

ly: mber tted	....	**IMPORTANT** **If you have any queries, please contact ISAC Secretariat:** ISAC@cprd.com

itle
rison of cardiovascular outcomes observed in patients with type 2 diabetes and those ted by the UKPDS Risk Engine

is protocol describe a purely observational study using CPRD data (this may include the review of nised free text)?

☒ No ☐

is protocol seek access to data held under the CPRD Data Linkage Scheme?

☒ No ☐

re seeking access to data held under the CPRD Data Linkage Scheme, please select the source(s) of linked ing requested.

pital Episode Statistics ☐ Cancer Registry Data ☐ MINAP
S Mortality Data ☐ Index of Multiple Deprivation/ Townsend Score
her Baby Link ☐ Other: (please specify)

re seeking access to data held under the CPRD Data Linkage Scheme, have you already discussed your with a member of the Research team?

☒ No* ☐

*e contact the CPRD Research Team on +44 (20) 3080 6383 or email kc@cprd.com to discuss your ments **before** submitting your application.*

is protocol involve requesting any additional information from GPs?

☐ No ☒

e indicate what will be required:
tion of questionnaires by the GP. Yes ☐ No ☐
n of anonymised records (e.g. hospital discharge summaries) Yes ☐ No ☐
please describe)

onnaire for completion by GPs needs to be approved by ISAC before being sent out for completion.

E ON ANSWERING QUESTIONS 2-3:
stions must be completed by all applicants. You should note the following:

have answered NO to question 2, you may need to seek separate ethics approval from an NHS rch Ethics Committee for this study. The ISAC will provide advice on whether this may be d.

have answered YES answered to question 2 above and you will be using data obtained from the Group at the MHRA (question 3), this study does not require separate ethics approval from an esearch Ethics Committee.

be using data obtained from EPIC, you will need to consult the data provider regarding their ents for obtaining ethics approval for the study.

NB: Answering YES to question 2 means that the answers to questions 8-10 should all be NO. If any of the answers below are YES please review your answer to question 2 as it should be NO.

7. Has this protocol been peer reviewed by another Committee?

Yes* ☐ No ☒

** Please state in your protocol the name of the reviewing Committee(s) and provide an outline of the review process and outcome.*

8. Does the study involve linking to patient *identifiable* data from other sources?

Yes ☐ No ☒

9. Does this study require contact with patients in order for them to complete a questionnaire?

Yes ☐ No ☒

10. Does this study require contact with patients in order to collect a sample?

Yes* ☐ No ☒

** Please state what will be collected*

11. Type of Study *(please tick one box below)*

Adverse Drug Reaction	☐	Drug Use	☐	Disease Epidemiology	☒
Pharmacoeconomic	☐	Drug Effectiveness	☐	Other	☐

12. Data source *(please tick one box below)*

CPRD :

Sponsor has on-line access	☒	Purchase of ad hoc dataset	☐
Commissioned study	☐		

Other ☐ *(please specify)*

13. Financial Sponsor of study

Pharmaceutical Industry *(please specify)*	☐	Academia*(please specify)*	☒ Cardiff University
Government / NHS *(please specify)*	☐	None	☐
Other *(please specify)*	☐		

14. This study is intended for:

Publication in peer reviewed journals	☒	Presentation at scientific conference	☐
Presentation at company/institutional meetings	☐	Other	

15. Principal Investigator (full name, job title, organisation & e-mail address for correspondence regarding this protocol)
Christian Bannister, MSc, Postgraduate Researcher, Cardiff University, bannisterca@cf.ac.uk

16. Affiliation (full address)
Cardiff University School of Medicine, Primary Care & Public Health, Cardiff Medicentre, Heath Park, Cardiff, CF14 4UJ

17. Type of Institution (please tick one box below)

Academia	☒	Research Service Provider	☐	Pharmaceutical Industry	☐
NHS	☐	Government Departments	☐	Others	☐

18. Experience/expertise available

Please complete the following questions to indicate the experience/expertise available within the team of researchers actively involved in the proposed research, including analysis of data and interpretation of results

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	☐	☐
1-3	☐	☐
> 3	☒	☒

	Yes	No
Is statistical expertise available within the research team?	☒	☐
If yes, please outline level of experience	*The research team has statistical expertise appropriate to use of large datasets for epidemiological research. Additional expertise can be sought from colleagues within Cardiff School or Medicine as required.*	
Is experience of handling large data sets (>1 million records) available within the research team?	☒	☐
If yes, please outline level of experience	*CP and CC and SJJ have extensive experience of using large routine NHS datasets including HES, GRPD and THIN.*	
Is UK primary care experience available within the research team?	☐	☐
If yes, please outline level of experience	*The team has experience of using primary care data form a variety of studies. Specific clinical assistance can be sought if required from colleagues from the department of Primary Care and Public Health within Cardiff School of Medicine*	

19. Other collaborators (if applicable: *please list names and affiliations of all collaborators*)
Dr. Chris Poole
Senior Lecturer in Evaluation of Medicines, Cardiff University
Sara Jenkins-Jones, MSc
Postgraduate Researcher, Cardiff University
Professor Craig Currie
Professor of Applied Pharmacoepidemiology, Cardiff University

20. Protocol's Author (if different from PI)

PROTOCOL CONTENT CHECKLIST

All protocols using CPRD data which are submitted for review by ISAC must contain information on the areas detailed in the instructions. If you do not feel that a specific area required by ISAC is relevant for your protocol, you will need to justify this decision to ISAC.

Applicants must complete the checklist below to confirm that the protocol being submitted includes all the areas required by ISAC, or to provide justification where a required area is not considered to be relevant for a specific protocol. Protocols will not be circulated to ISAC for review until the checklist has been completed by the applicant.

Please note, your protocol will be returned to you if you do not complete this checklist, or if you answer 'no' and fail to include justification for the omission of any required area.

	Included in protocol?		
Required area	**Yes**	**No**	**If no, reason for omission**
Lay Summary (max.200 words)	☒	☐	
Background	☒	☐	
Objective, specific aims and rationale	☒	☐	
Study Type *Hypothesis Generating* *Hypothesis Testing*	 ☒ ☒	 ☐ ☐	
Study Design	☒	☐	
Sample size/ power calculation (Please provide detailed justification of sample size in the protocol)	☒	☐	
Study population (including estimate of expected number of relevant patients in the CPRD)	☒	☐	
Selection of comparison group(s) or controls	☐	☒	**N/A to this study**
Exposures, outcomes and covariates	☒	☐	
Data/ Statistical Analysis *Hypothesis Generating* *Hypothesis Testing*	 ☒ ☒	 ☐ ☐	
Patient/ user group involvement †	☐	☒	**N/A to this study**
Limitations of the study design, data sources and analytic methods	☒	☐	
Plans for disseminating and communicating study results	☒	☐	

† ***It is expected that many studies will benefit from the involvement of patient or user groups in their planning and refinement, and/or in the interpretation of the results and plans for further work. This is particularly, but not exclusively true of studies with interests in the impact on quality of life. Please indicate whether or not you intend to engage patients in any of the ways mentioned above.***

ISAC strongly recommends that researchers using CPRD consider registering as a NRR data provider in order that others engaged in research within the UK can be made aware of current works. The ***National Research Register (NRR)*** *is a register of ongoing and recently completed research projects funded by, or of interest to, the United Kingdom's National Health Service. Information on the NRR is available on www.nrr.nhs.uk .*

Please Note: Registration with the NRR is entirely voluntary and will not replace information on ISAC approved protocols that are published in summary minutes or in the ISAC annual report.

GPRD ISAC protocol

Comparison of cardiovascular outcomes observed in patients with type 2 diabetes and those predicted by the UKPDS Risk Engine

Christian Bannister, MSc
Postgraduate Researcher, Cardiff University

Dr. Chris Poole
Senior Lecturer in Evaluation of Medicines, Cardiff University

Sara Jenkins-Jones, MSc
Postgraduate Researcher, Cardiff University

Professor Craig Currie
Professor of Applied Pharmacoepidemiology, Cardiff University

Submitted on behalf of Cardiff University

TABLE OF CONTENTS

Lay summary

The UK Prospective Diabetes Study (UKPDS) was a clinical trial designed to show the long-term benefits of controlling both blood sugar and blood pressure in individuals newly diagnosed with type 2 diabetes (T2DM). The UKPDS Risk Engine was developed from the trial data and calculates the absolute risk of coronary heart disease and stroke for individuals diagnosed with T2DM. The UKPDS Risk Engine is widely utilised in the economic evaluation of various T2DM-related interventions across the world as well as in the UK.

A growing number of trials have failed to demonstrate the causal relationship between lower blood glucose and improved outcomes, some even suggesting that there is a harmful effect of lowering blood sugar too aggressively. Therefore the benefits of improved glycaemic control predicted by models such as the UKPDS Risk Engine may be overestimated.

The purpose of this epidemiological study is to compare observed cardiovascular outcomes with those predicted by the UKPDS Risk Engine in a representative population of people with T2DM managed according to routine standard of care. This study of real-life data should give a better understanding of the applicability of the UKPDS Risk Engine to the general population of people with T2DM.

Introduction

Diabetes is on the rise, in the UK and around the world[1-7]. Forecasting models have shown that the prevalence of diabetes is steadily increasing, and that diabetes is not a localised chronic condition[8-12]. Major contributors to this increase in prevalence are obesity and an ageing population, both of which increase risk of type 2 diabetes (T2DM). Understanding the costs and effectiveness of healthcare delivery in diabetes is of clear importance to health services.

There are many different predictive risk models for diabetes used in health economics, such as the UKPDS Risk Engine[13], the UKPDS Outcomes Model[14] and the Center for Outcomes Research Diabetes Model[15]. All these models are based on epidemiological data derived from the UKPDS[23]. As such, while assumptions may hold true in a clinical setting with a specific population of individuals who have specific risk factors used in the UKPDS, these models may not accurately reflect the actual incidence, prevalence, mortality and costs related to the late complications of diabetes in the general population today or after a given duration[16].

Moreover, models derived from UKPDS epidemiological data[13-15] include estimates of reduced mortality and cardiovascular events associated with a reduction in HbA1c for patients with T2DM. However, a growing number of large randomised, controlled trials have failed to demonstrate a causal relationship between clinically important reductions in HbA1c and improved outcomes[16-20].

The recent Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial[21] and retrospective cohort study assessing survival as a function of HbA_{1c}[22] even suggests potential harm associated with aggressive glycaemic control in this patient population. The benefits of

improved glycaemic control for people with T2DM may therefore be overestimated in these models.

Another important factor is the manner in which the UKPDS Risk Engine[13] is used in health economics. The risk engine was designed to predict the future risk of specific cardiovascular events in patients newly diagnosed with T2DM over a specified time horizon. However, the model is often used at other stages of the care pathway, not just at the beginning as intended, with the assumption that the model performs the same at any level of treatment intensification. The validity of this assumption is far reaching as many T2DM interventions have been evaluated in this way for use by the NHS and other health services worldwide.

A review of previous work carried to externally validate the UKPDS risk engine revealed that the risk engine has been previously validated on prospective cohorts; 10,137 from Norfolk[23] ; 428 from Poole[24]; 1,622 from Germany and the Netherlands[25]; 1,482 from the Netherlands[26] and 7,067 from China[27]. However none of these studies have validated the UKPDS risk engine on a UK-wide prospective cohort, nor have any studies evaluated its performance at different stages of the diabetes care pathway.

In the light of new research that challenges our understanding of the impact of glycaemic control on long-term outcomes in T2DM, this study requests approval to use GPRD data in evaluating the appropriateness of the UKPDS Risk Engine and the assumption that the Risk Engine can be used at any stage of the T2DM treatment pathway.

Objective

The purpose of the proposed study is to characterise the appropriateness of the UKPDS Risk Engine for the prediction of cardiovascular events for individuals with T2DM from the general population.

Methods

Data source

GPRD

Study type

The study will be hypothesis testing with regard to cardiovascular risk in patients with T2DM.

Study design

The study will use a retrospective cohort design.

Study population and cohorts

The aim of the study is not to replicate the UKPDS, but to access the appropriateness of using the UKPDS Risk Engine (developed using data from the UKPDS) in the UK T2DM population. We achieve this by adopting selection criteria to develop a cohort of incident cases of T2DM, with no recent history of myocardial infarction (MI), angina or heart failure from GPRD and test the performance of the risk engine on this cohort.

The study comprises of a single cohort of incident cases of type 2 diabetes registered with practices between 1991 and 2011. A wash-in of 365 days has been applied to exclude non-incident T2DM patients (see table 1, item 5). Patients will be selected who have had a diagnosis of type 2 diabetes with no frank diagnoses of diabetes secondary to other causes e.g. gestational or iatrogenic. Subjects will be classified as type 2 diabetes on the basis of frank differential diagnosis. More specifically, patients will be classified as having T2DM if they have a diagnosis of diabetes and one or more of the following:

1. More than one diagnostic record exclusively for type 2 diabetes OR
2. Prescription of two or more differing classes of OAD OR
3. A diagnostic code indicative of T2DM (regardless of conflicting diagnoses of type 1 or non-specific diabetes) plus a prescription for an OAD.

Patients diagnosed before the age of 35 with no OAD and a prescription for insulin were classified as type 1 and hence excluded. In the absence of such a diagnosis, a prescription history including exposure to oral hypoglycaemic agents (assignment to T2DM if >6 separate OHA prescriptions) where the age at incident diabetes event is greater than 40 years of age.

Patients will be selected that have linkage with Hospital Episode Statistics (HES) data and Office for National Statistics (ONS) mortality data. This will provide the ethnicity and cause of death information required for this study as well as more accurate ascertainment of cardiovascular events. This will possibly result in some data loss (see table 1) as the data that is eligible for HES and/or ONS linkage is a subset of the data available through GPRD.

Stage	Description	Lost	Remaining
1	GPRD patients with T2DM of status 'accept' (GPRD quality indicator)	N/A	490,084
2	Omit patients in 1 with unknown dates of first diagnosis or first prescription	14,096	475,988
3	Omit patients in 2 where year of DM presentation* > year of birth or > 2011	1	475,987
4	Omit patients in 3 where year of DM presentation* < 1991	44,462	431,525
5	Omit patients in 4 where wash-in from registration to DM presentation* < 365 days	106,406	325,119
6	Omit patients in 5 where sex is neither male nor female	24	325,095
7	Omit patients in 6 where age at DM presentation* < 21	4,838	320,257
9	Omit patients in 7 not having smoking status at baseline (DM presentation*)	8,897	311,360
8	Omit patients in 7 who are ineligible for HES linkage	180,205	131,155
10	Omit patients in 9 not having ethnicity of types 1, 2, or 3 (see 'Model Covariates' section)	26,512	104,643
11	Omit patients in 10 who are ineligible for ONS Mortality linkage (and hence cause of death)	778	103,865
**DM Presentation refers to the earlier of two dates; date of first DM diagnosis or date of first prescription of DM drugs*			

Table 1: Sample data loss from inclusion/exclusion criteria

Sample size and power calculation

The UKPDS risk engine was developed on 4,540 patients from the UKPDS and is conservatively assumed as the minimum sample size for this study. This minimum number of subjects is easily achievable, where we estimate that >100,000 suitable subjects with incident of T2DM are available in GPRD; even with 'acceptable quality status' (see table 1).

Outcomes

The primary outcomes are the four cardiovascular events predicted by the UKPDS Risk Engine are coronary heart disease (CHD), fatal CHD, stoke and fatal stroke. This study proposes to use the same endpoints to ensure that we are predicting the same events. Using ICD-9and ICD-10, Table 2 below details the exact coding of the endpoints used in UKPDS[28] risk engine that are proposed for this study. For hard endpoints of stroke and CHD we will be reliant on data from HES and ONS. For HES we will consider patients admitted as emergencies with a primary diagnosis of the relevant event. For ONS we will consider deaths with a primary or contributory cause of death of stoke or CHD.

Endpoint	**ICD -10 Codes**	**ICD-9 Code**
CHD	I21-I25, I46.1	410-414.9 798.9 (fatal or non-fatal MI and Sudden death)
Fatal CHD	I21-I25, I46.1	410-414.9 798.9 (fatal MI or Sudden death)
Stroke	I60-I69	430-438.9 (fatal or non-fatal Stroke)
Fatal Stroke	I60-I69	430-438.9 (fatal Stroke)

Table 2: ICD-9 and ICD-10 coding of cardiovascular endpoints used in UKPDS and proposed for study.

CHD is defined as the occurrence of fatal or non-fatal myocardial infarction (MI) or sudden death.[28] In patients with multiple CHD events, only the first event is considered in this study. Stroke is defined as a neurological deficit with symptoms or signs lasting 1 month or more.[28] No distinction was made between ischemic, embolic, and hemorrhagic strokes. In patients with multiple strokes, only the first stroke is considered here. Death from causes other than the defined outcomes of interest will be treated as censored.

Model covariates

Covariates are those required by the UKPDS Risk Engine as inputs, as follows:

AGE	Age in years (over 20 years) at diagnosis
SEX	F for female; M for male
ETHNIC	1 for Caucasian, 2 for Afro-Caribbean; 3 for Asian-Indian
SMOK	0 for never, 1 for past, 2 for current smoker of tobacco in any form, at index of each care pathway
DUR	Duration of diabetes (positive integer)
AF	Presence of Atrial Fibrillation (Y for yes, N for no)

SBP	Systolic blood pressure (mmHg), mean of values for years 1 and 2
A1C	HbA1c (%), mean of values for years 1 and 2
TC	Total cholesterol, mean of values for years 1 and 2 (mmol/l)
HDL	HDL cholesterol, mean of values for years 1 and 2 (mmol/l)
H	No of readings for HbA1c used in calculating mean value used
BP	No of readings for SysBP used in calculating mean value used
CHOL	No readings TC/HDL used in calculating mean value used

To improve model stability HbA1c, SysBP, TC and HDL will be taken as mean values one year apart

Data/statistical analysis

The UKPDS Risk Engine will be evaluated with a series of diagnostic plots, comparing survival probabilities for the study population calculated by the UKPDS Risk Engine with survival probabilities for the study population calculated by non- parametric methods. A version of the UKPDS Risk Engine has been obtained for this purpose, with the permission of the Diabetes Trails Unit at the University of Oxford. The non-parametric method proposed is the life-table method with one-year intervals, which also provides 95% confidence intervals[29]. The life-table method, also known as *Actuarial Estimate of survivor function* is a technique for estimating the survival function *S(t)* of censored data. The life-table method is analogous to the Kaplan-Meier estimate of the survival function, but with the life-table method intervals can predefined and uniform (e.g. 1-year), whereas with Kaplan-Meier time intervals are defined by event times. A more detailed explanation is outlined in appendix A. This methodology is the same used in the development and evaluation of the UKPDS Risk Engine[13] and should therefore aid comparison. Declining secular trends in diabetes mortality and all-cause mortality in the general population shall be addressed by performing a sensitivity analysis using 5-year periods. Where appropriate missing data shall be handled using multiple imputation. Multiple imputation is a powerful technique that offers substantial improvements over the biased and flawed value replacement approaches based on complete cases or cases matched for age and sex[30-31]. It involves creating multiple copies of the data and imputing the missing values with sensible values randomly selected from their predicted distribution. We propose to use the Multivariate Imputation by Chained Equations (MICE) approach.

In order to evaluate the performance of the UKPDS Risk Engine at various stages of the diabetes care pathway, four levels of glycaemia treatment levels have been defined:

1. Diet and lifestyle modification
2. Metformin monotherapy
3. Metformin in combination with other oral hypoglycaemic agents (OHA)
4. Insulin–based therapy (with or without OHAs)

The UKPDS Risk Engine will be evaluated using all available patients in the study population as they enter each of the pre-defined treatment pathways, this will be carried out for each of the four outcomes. The behaviour of the UKPDS Risk Engine at each of these treatment stages will be compared and contrasted with the parametric life-table method using diagnostic plots as described above.

Limitations of the study design, data sources and analytic methods

This study has a number of limitations. GPRD collates data from routine practice, thus there are missing data, there are coding imperfections, and there is no standardisation of measures such as HbA1c. The normal ranges for HbA1c do vary between biochemical test centers (unless test is specifically reported as DCCT−aligned), and measurements are taken with varying periodicity. We consider these limitations may introduce noise into the study but have no reason to suspect they will introduce bias. Duration of diabetes is an important covariate but one which must be treated with caution.

Plans for disseminating and communicating study results

Findings from this study will be disseminated through scientific meetings and peer-reviewed manuscript(s).

References

1 Jönsson B. Revealing the cost of Type II diabetes in Europe. Diabetologia. 2002. 45 (7), S5-S12

2 Venkat Narayan K, Gregg E, Fagot-Campagna A, Engelgau M, Vinicor F. Diabetes — a common, growing, serious, costly, and potentially preventable public health problem, Diabetes Research and Clinical Practice, 50(2), 2000, S77-S84

3 Bagust A, Hopkinson P, Maslove L, Currie C. The projected health care burden of Type 2 diabetes in the UK from 2000 to 2060. Diabetic Medicine. 2002. 19(s4) 1–5

4 Canadian Diabetes Association. The prevalence and costs of diabetes. http://www.diabetes.ca/diabetes-and-you/what/prevalence/. Accessed Feb 2011.

5 Chodick G, Heymann A, Wood F, Kokia E. The direct medical costs of diabetes in Israel. Eur J Health Econ. 2005;6:166-171.

6 Rubin R, Altman W, Mendelson D. Health care expenditures for people with diabetes mellitus, 1992. J Clin Endocrinol Metab. 1994;78:809A-809F.

7 Williams R, Van Gall L, Lucioni C. Assessing the impact of complications on the costs of type II diabetes. Diabetologia. 2002;45:S13-S17.

8 King H, Aubert R, German W. Global burden of diabetes. Diabetes Care. 1998;21:1414-1431.

9 Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes; estimates from the year 2000 and projections for 2030. Diabetes Care. 2004;27:1047-1053.

10 Mainous A, Baker R, Koopman R, et al. Impact of population at risk of diabetes on projection of diabetes burden in the United States: an epidemic on the way. Diabetologia. 2007;50:934-950.

11 Boyle J, Honeycutt A, Narayan K, et al. Projection of diabetes burden through 2050. Diabetes Care. 2001;24:1936-1940.

12 Ohinmaa A, Jacobs P, Simpson S, Johnson J. The projection of prevalence and cost of diabetes in Canada: 2000 to 2016. Can J Diabetes. 2004;28(2):116-123.

13 Stevens RJ, Kothari V, Adler AI, Stratton IM. The UKPDS Risk Engine: a model for the risk of coronary heart disease in type 2 diabetes (UKPDS 56). Clinical Science 2001;101:671-679

14 Clarke P, Gray A, Briggs A, et al. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no.68). Diabetologia. 2004;47:1747-1759.

15 Palmer A, Roze S, Valentine W, et al. The CORE Diabetes Model: projecting long-term clinical outcomes costs and cost-effectiveness of intervention in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. Curr Med Res Opin. 2004;20:S5-S26.

16 Lau R, Ohinmaa A, Johnson JA (2011) Predicting the future burden of diabetes in Alberta from 2008 to 2035. *Canadian Journal of Diabetes* 35: 274-281

17 UK Prospective Diabetes Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes. UKPDS 33. Lancet. 1998;352:837-853.

18 Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med. 2008;358:2545-2559.

19 VADT Investigators. Glucose control and vascular complications in veterans with type 2 diabetes. N Engl J Med. 2009; 360:129-139.

20 ADVANCE Collaborative Group. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. N Engl J Med. 2008;358:2560-2572.

21 Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med. 2008;358:2545-2559.

22 Currie CJ, Peters JR, Tynan A, Evans M, Heine RJ, Bracco OL, Zagar T, Poole CD. Survival as a function of HbA1c in people with type 2 diabetes: a retrospective cohort study. LANCET 375(9713):481-489 2010

23 Rebecca K. Simmons, Ruth L. Coleman, Hermione C. Price, Rury R. Holman, Kay-Tee Khaw, N. J. W. and S. J. G. (2009). Performance of the UK prospective diabetes study risk engine and the Framingham risk equations in estimating cardiovascular disease in the EPIC-Norfolk cohort. *Diabetes Care*, *32*(4). doi:10.2337/dc08-1918.

24 Guzder, R. N., Gatling, W., Mullee, M. a, Mehta, R. L., & Byrne, C. D. (2005). Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed Type 2 diabetes: results from a United Kingdom study. *Diabetic medicine*, *22*(5), 554–62. doi:10.1111/j.1464-5491.2005.01494.x

25 van Dieren, S., Peelen, L. M., Nöthlings, U., van der Schouw, Y. T., Rutten, G. E. H. M., Spijkerman, a M. W., van der A, D. L., et al. (2011). External validation of the UK Prospective Diabetes Study (UKPDS) risk engine in patients with type 2 diabetes. *Diabetologia*, *54*(2), 264–70. doi:10.1007/s00125-010-1960-0

26 van der Heijden, A. a W. a, Ortegon, M. M., Niessen, L. W., Nijpels, G., & Dekker, J. M. (2009). Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. *Diabetes care*, *32*(11), 2094–8. doi:10.2337/dc09-0745

27 Yang, X., So, W.-Y., Kong, A. P. S., Ma, R. C. W., Ko, G. T. C., Ho, C.-S., Lam, C. W. K., et al. (2008). Development and validation of a total coronary heart disease risk score in type 2 diabetes mellitus. *The American journal of cardiology*, *101*(5), 596–601. doi:10.1016/j.amjcard.2007.10.019

28 UKPDS Group. UK Prospective Diabetes Study VIII: study design, progress and performance. Diabetologia. 1991;34:877–890.

29 Collett, D. (2003) Modelling Survival Data in Medical Research, pp 16-19, Chapman & Hall, London

30 Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing withmissing predictor values when applying clinical prediction models. Clin Chem 2009;55:994-1001.

31 Schafer JL. Multiple imputation: a primer. StatMethodsMed Res 1999;8:3-15.

Appendix A: Life-Table Method

The *life-table estimate of the survivor function*, also known as the *Actuarial estimate of the survivor function*, is obtained by first dividing the period of observation into a series of time intervals. These intervals need not necessarily be of equal length but often are.
Suppose that the jth of m such intervals, j=1, 2,…, m, extends from time t'_j to t'_{j+1}, and let d_j and c_j denote the number of deaths and number of censored survival times, respectively, in this time interval. Also let n_j be the number of individuals who are alive, and therefore at risk of death, at the start of the jth interval. We now make the assumption that the censored survival times occur uniformly throughout the jth interval, so that the average number of individuals who are at risk during this interval is

$$n'_j = \frac{n_j - c_j}{2}$$

This assumption is sometimes known as the actuarial assumption.

In the jth interval, the probability of death can be estimated by d_j/n'_j, so that the corresponding survival probability is $(n'_j-d_j)/n'_j$. Now consider the probability that an individual survives beyond time t'_k, k=1, 2, …, m, that is, until some time after the start of the kth interval. This will be the product of the probabilities that an individual survives beyond the start of kth interval and through each of the k-1 preceding intervals, and so the life-table estimate of the survivor function is given by

$$S^*(t) = \prod_{j=1}^{k} \left(\frac{n'_j - d_j}{n'_j} \right)$$

For $t'_k <= t < t'_{k+1}$, k=1, 2. …, m. A graphical estimate of the survivor function will then be a step-function with constant values of the function in each time interval.

Appendix B

Run statistics: SMART experiments

The full range of run statistics for the 25 GP runs in the SMART experiments in chapter 6

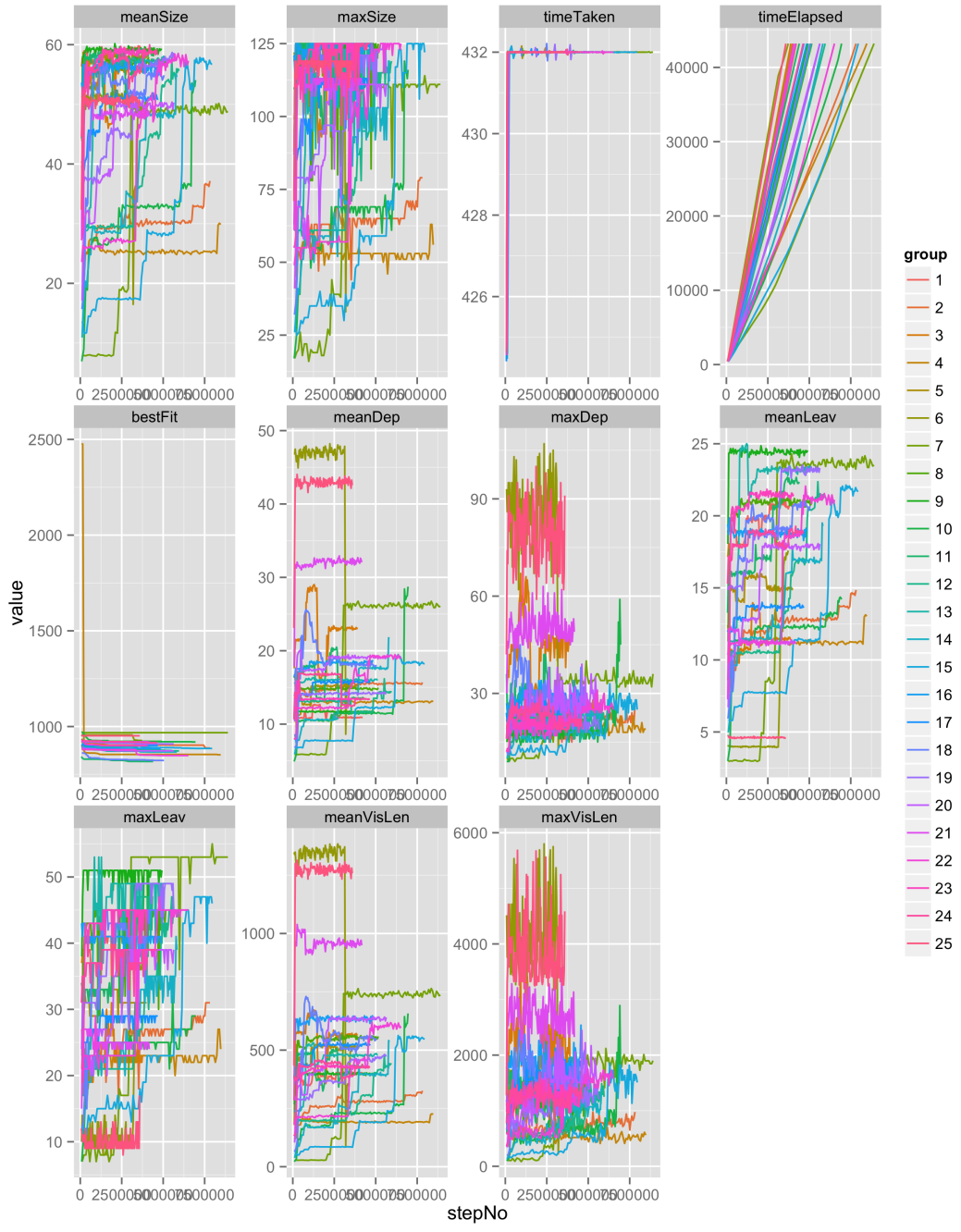
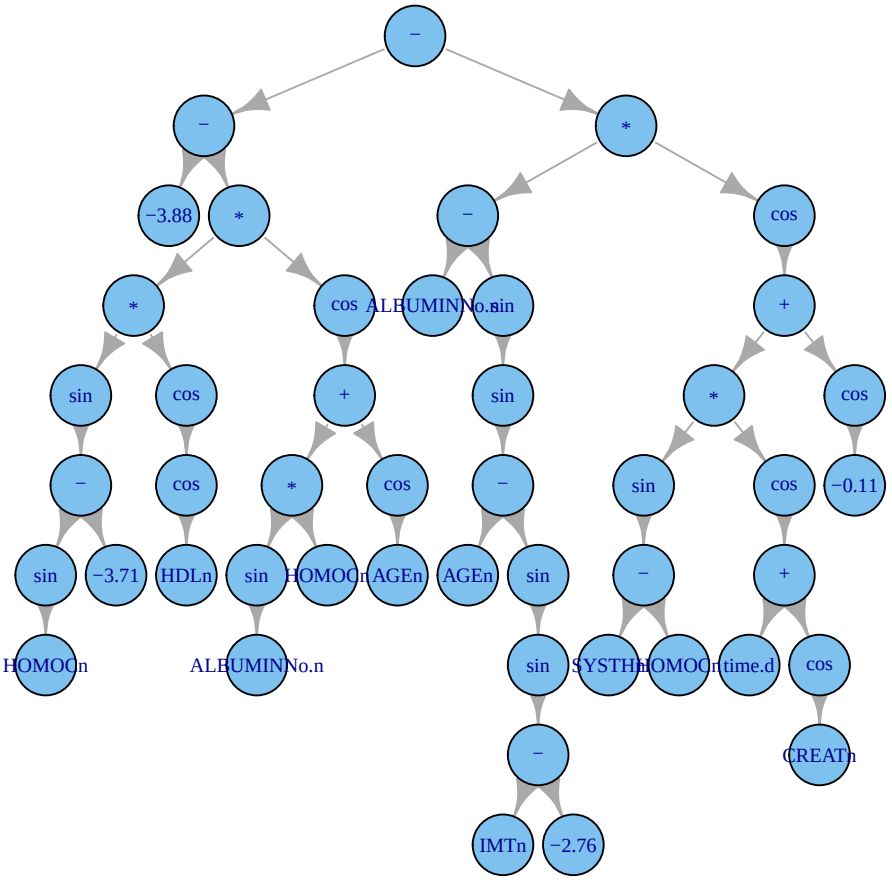


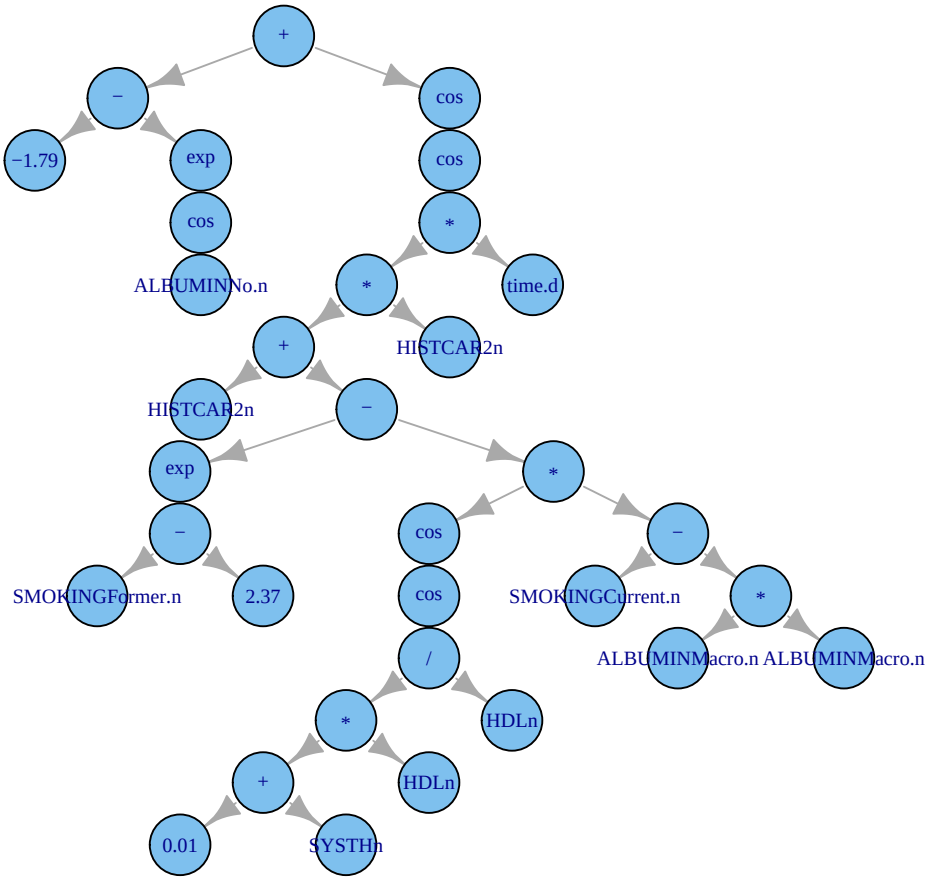
Figure B.1: The full range of runs statisitcs for the 25 SSOGP runs in the SMART experiments in chapter 6.

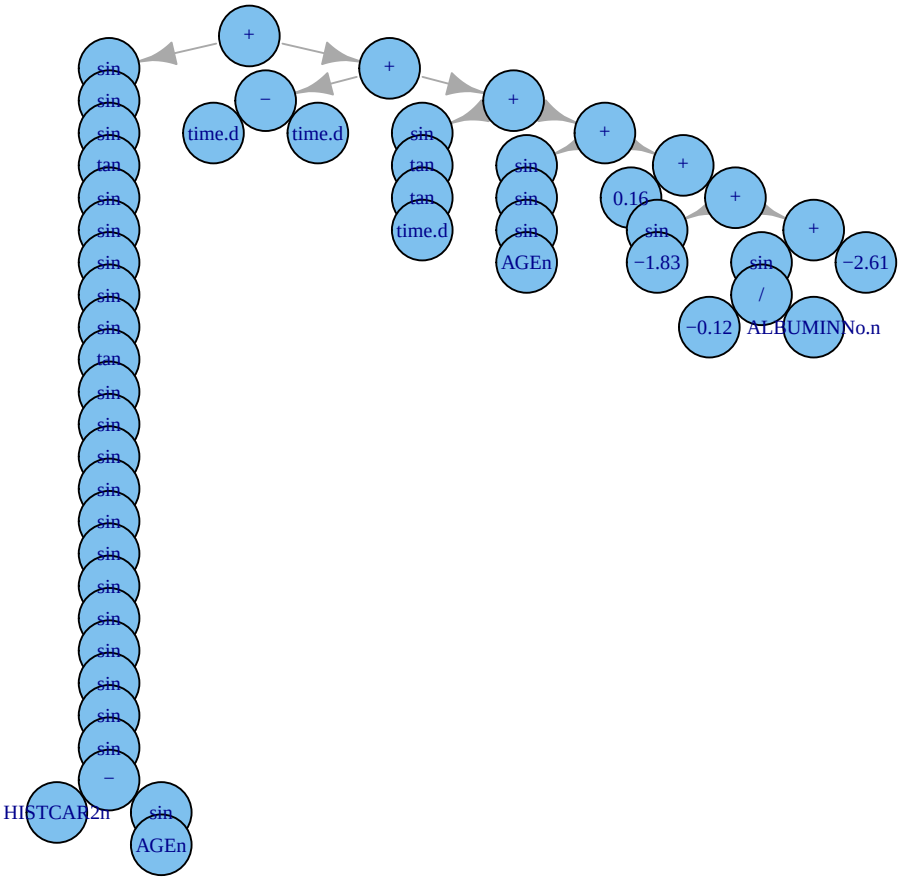
Appendix C

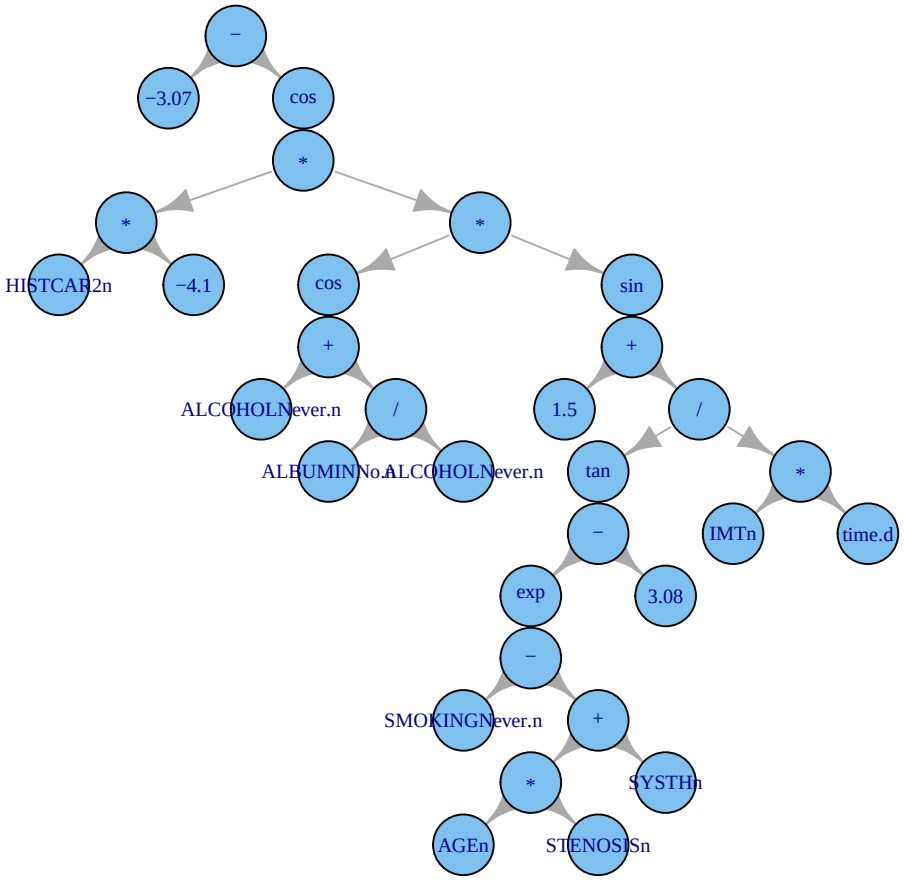
Final Models: SMART experiments

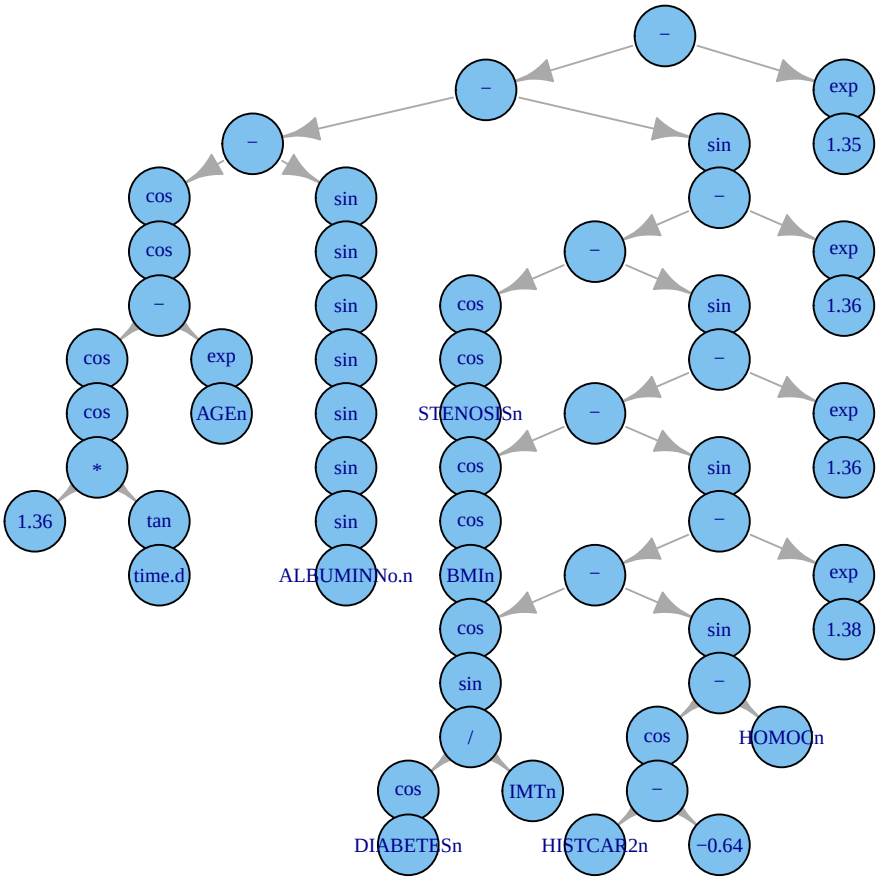
The final 25 models developed by SSOGP the SMART experiments in chapter 6, presented as a binary trees

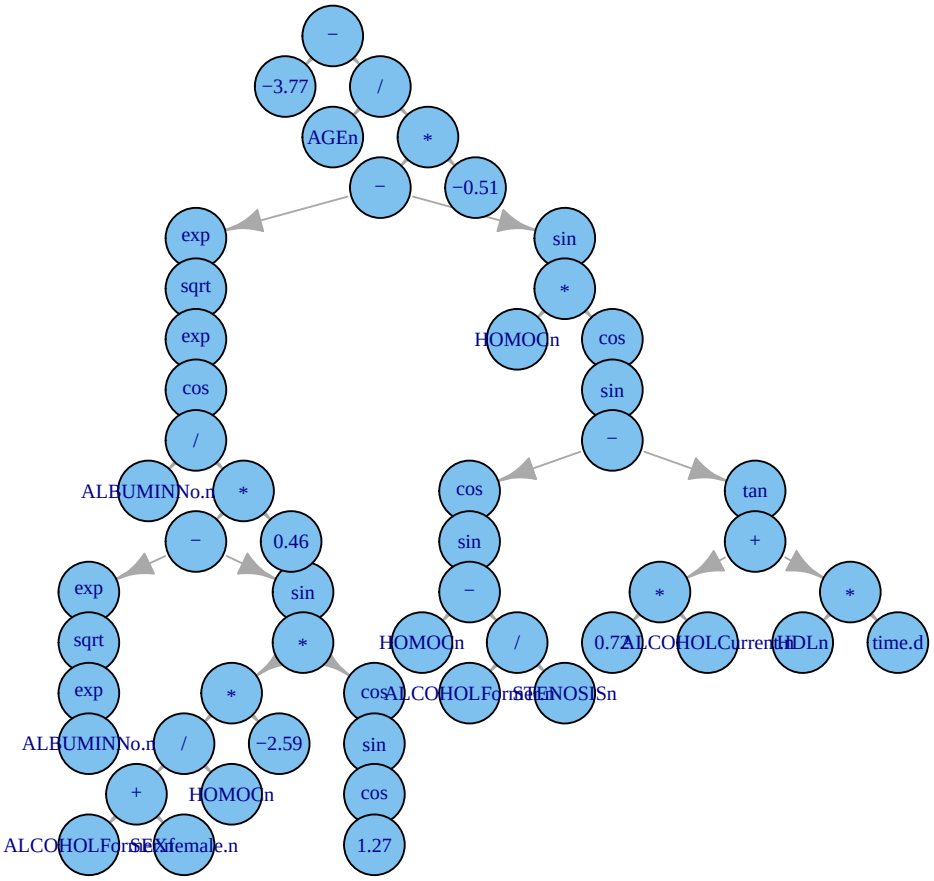

−
−
*
−3.88
*
−
cos
*
cos
ALBUMINNo.n
sin
+
sin
cos
+
sin
*
cos
−
cos
*
cos
−
sin
cos
−0.11
sin
−3.71
HDLn
sin
HOMOCn
AGEn
AGEn
sin
sin
−
+
HOMOCn
ALBUMINNo.n
sin
SYSTHn
HOMOCn
time.d
cos
−
CREATn
IMTn
−2.76

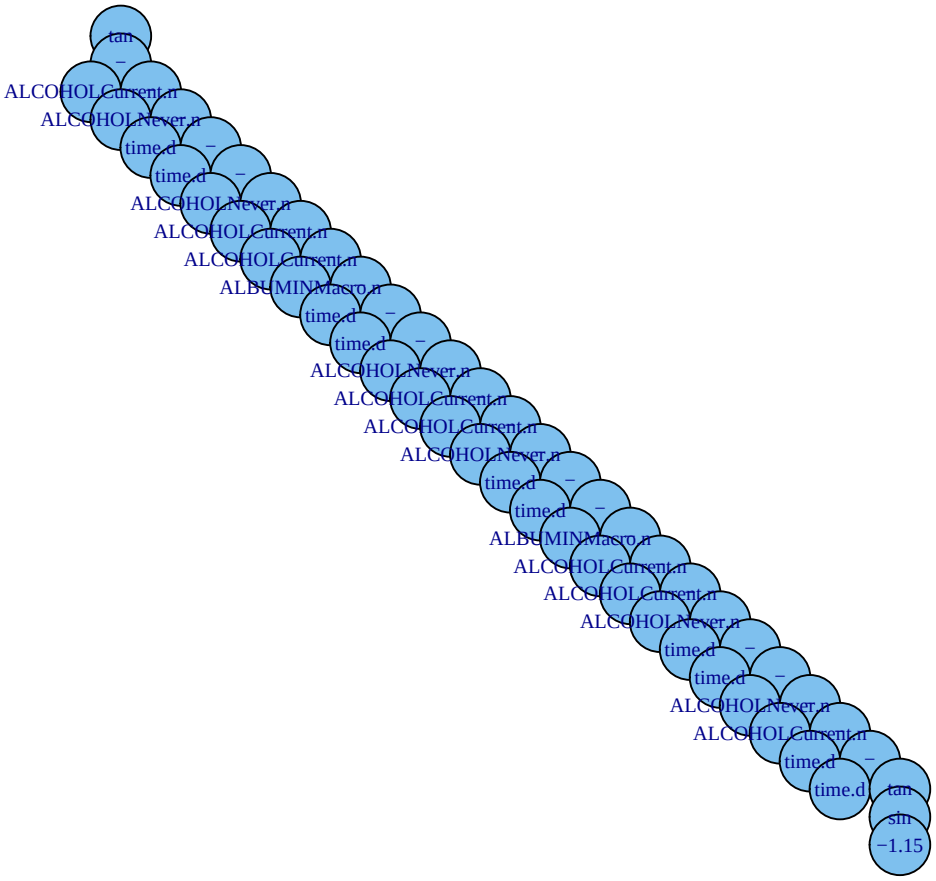

+
−
cos
−1.79
exp
cos
cos
*
ALBUMINNo.n
*
time.d
+
HISTCAR2n
HISTCAR2n
−
exp
*
−
cos
−
SMOKINGFormer.n
2.37
cos
SMOKINGCurrent.n
*
/
ALBUMINMacro.n
ALBUMINMacro.n
*
HDLn
+
HDLn
0.01
SYSTHn

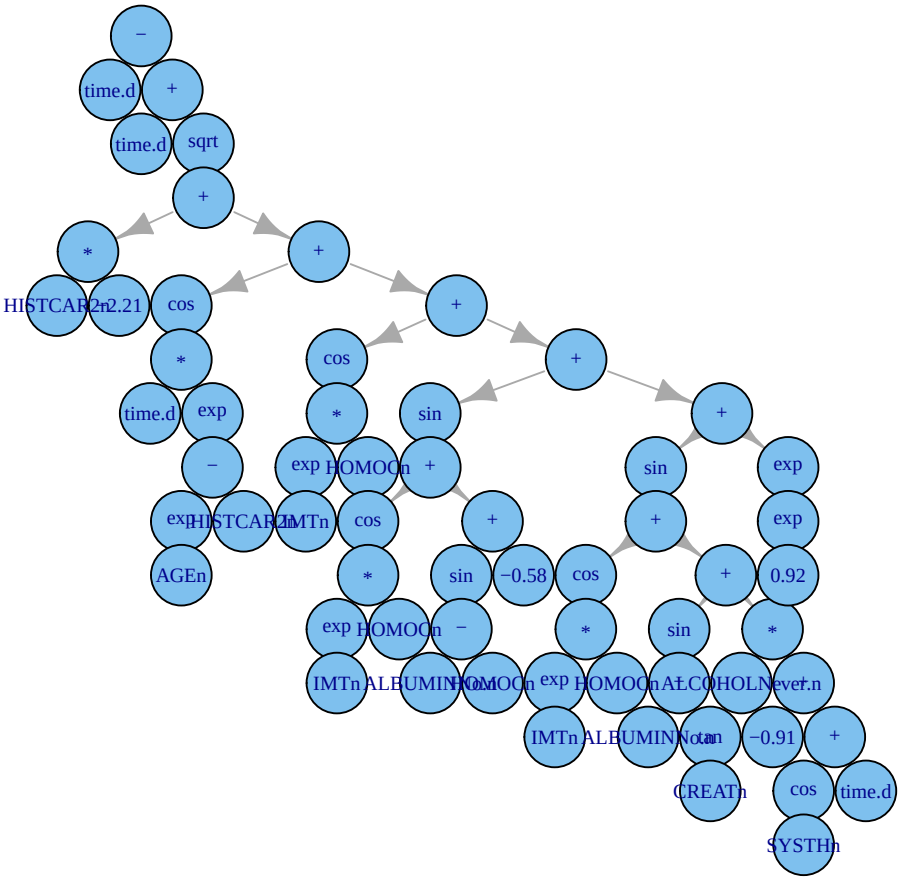

+
sin
sin
sin
tan
sin
sin
sin
sin
sin
tan
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
−
HISTCAH2n
sin
AGEn
+
−
time.d
time.d
+
sin
tan
tan
time.d
+
sin
sin
sin
AGEn
+
0.16
+
sin
−1.83
+
sin
/
−0.12
ALBUMINNo.n
+
−2.61

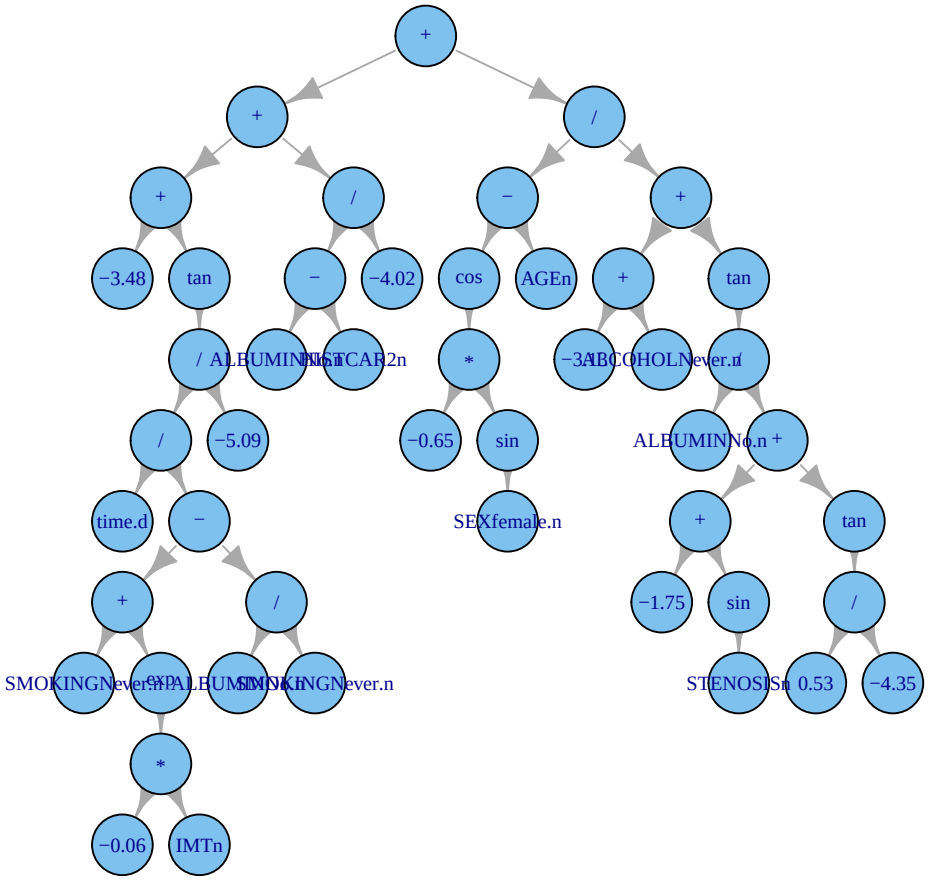

−
−3.07
cos
*
*
*
HISTCAR2n
−4.1
cos
sin
+
+
ALCOHOLNever.n
/
1.5
/
ALBUMINNo.
ALCOHOLNever.n
tan
*
−
IMTn
time.d
exp
3.08
−
SMOKINGNever.n
+
*
SYSTHn
AGEn
STENOSISn

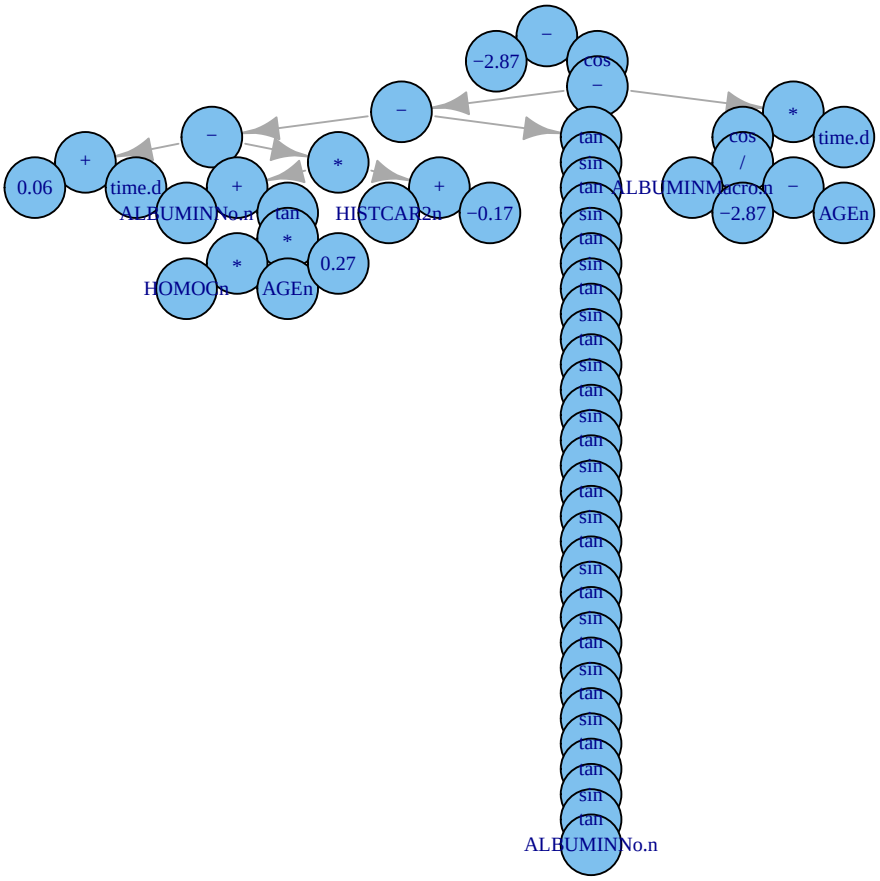

−
−
exp
1.35
−
sin
−
exp
1.36
−
sin
−
exp
1.36
−
sin
−
exp
1.38
−
sin
−
HOMOCn
cos
−
−0.64
HISTCAR2n
cos
IMTn
/
sin
cos
BMIn
cos
cos
STENOSISn
cos
cos
−
cos
exp
AGEn
cos
cos
*
1.36
tan
time.d
−
sin
sin
sin
sin
sin
sin
sin
ALBUMINNo.n
cos
DIABETESn

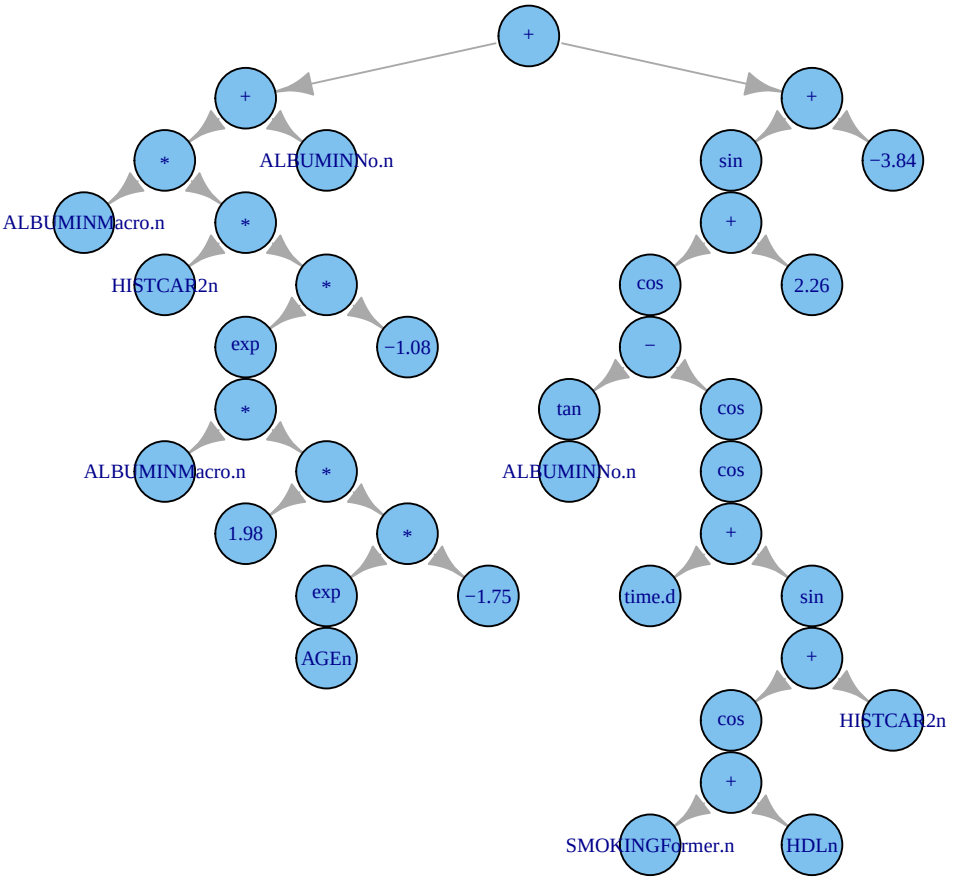

−
−3.77
/
AGEn
*
−
−0.51
exp
sqrt
exp
cos
/
ALBUMINNo.n
*
−
0.46
exp
sqrt
exp
ALBUMINNo.n
/
−2.59
+
HOMOCn
ALCOHOLFormerSEXfemale.n
sin
*
cos
sin
cos
1.27
sin
*
HOMOCn
cos
sin
−
HOMOCn
/
ALCOHOLFormerSEXfemale.n
tan
+
*
0.72
ALCOHOLCurrentHDLn
*
time.d

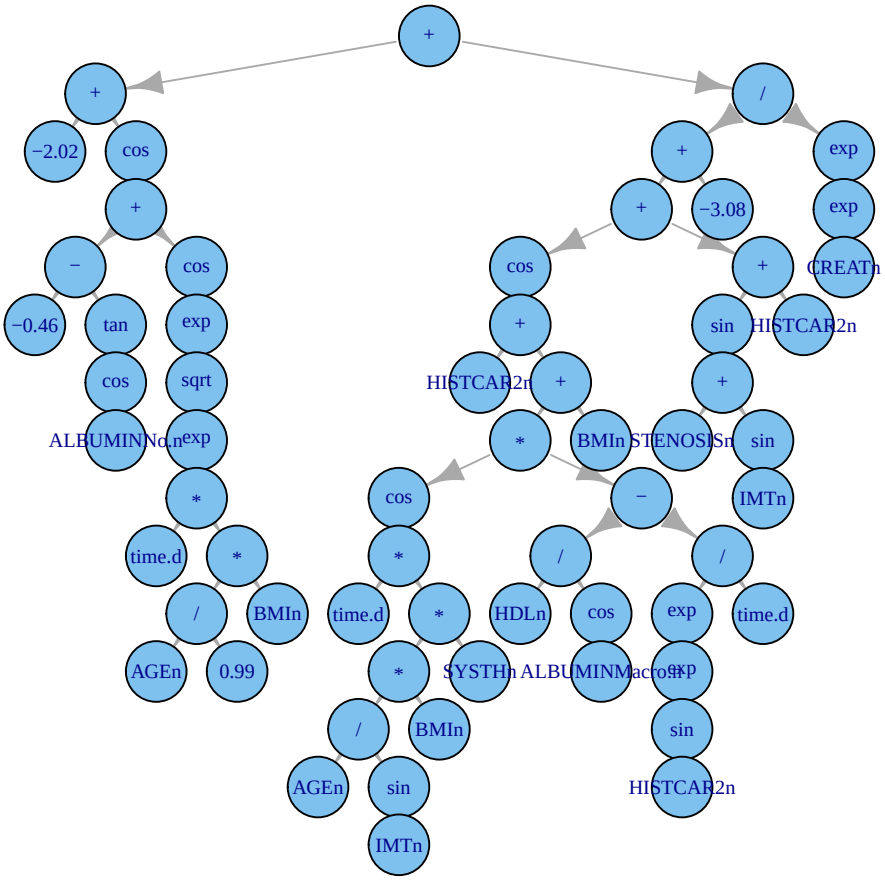

tan
−
ALCOHOLCurrent.n
ALCOHOLNever.n
time.d
−
time.d
−
ALCOHOLNever.n
ALCOHOLCurrent.n
ALCOHOLCurrent.n
ALBUMINMacro.n
time.d
−
time.d
−
ALCOHOLNever.n
ALCOHOLCurrent.n
ALCOHOLCurrent.n
ALCOHOLNever.n
time.d
−
time.d
−
ALBUMINMacro.n
ALCOHOLCurrent.n
ALCOHOLCurrent.n
ALCOHOLNever.n
time.d
−
time.d
−
ALCOHOLNever.n
ALCOHOLCurrent.n
time.d
−
time.d
tan
sin
−1.15

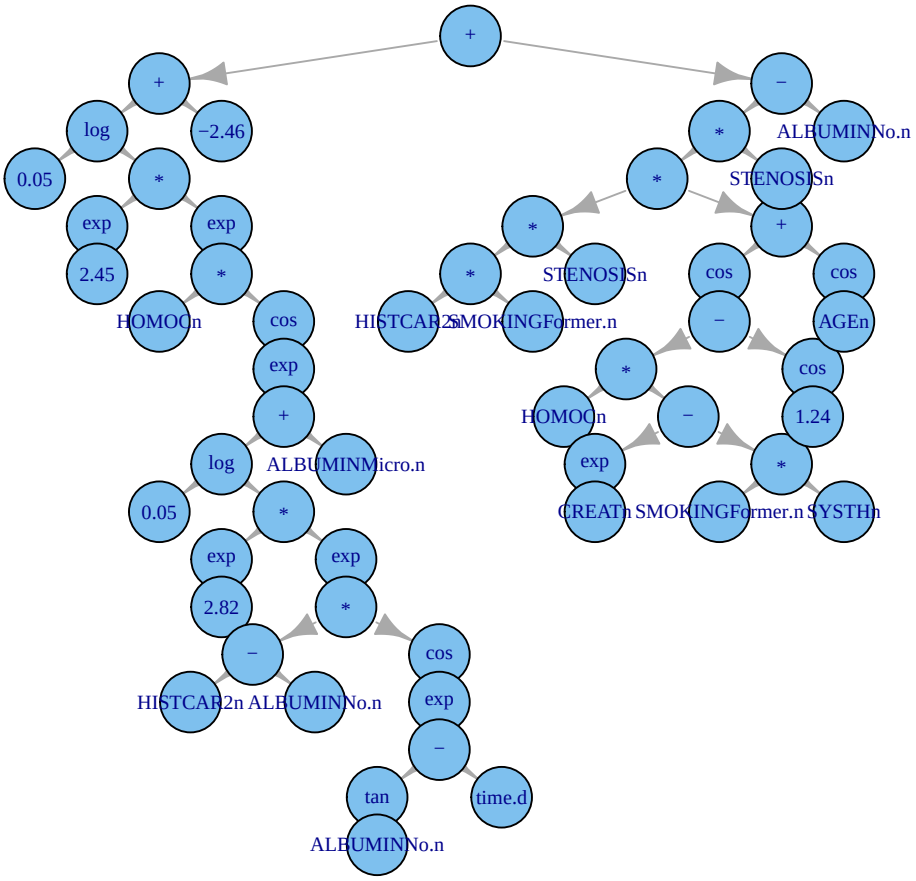

−
time.d
+
time.d
sqrt
+
*
+
HISTCAR2n
2.21
cos
+
*
cos
+
time.d
exp
*
sin
+
−
exp
HOMOCn
+
sin
exp
exp
HISTCAR2n
IMTn
cos
+
+
exp
AGEn
*
sin
−0.58
cos
+
0.92
exp
HOMOCn
−
*
sin
*
IMTn
ALBUMINn
HOMOCn
exp
HOMOCn
ALCOHOLn
ever.n
IMTn
ALBUMINn
tan
−0.91
+
CREATn
cos
time.d
SYSTHn

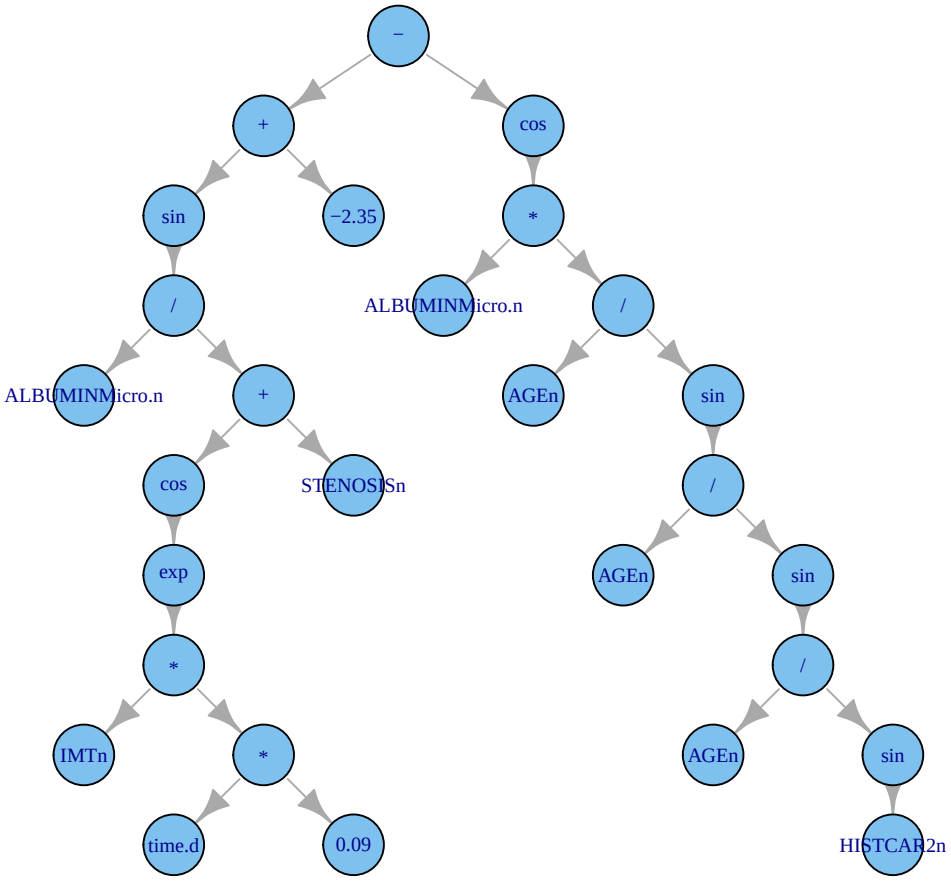

+
+
/
+
/
−
+
−3.48
tan
−
−4.02
cos
AGEn
+
tan
/ ALBUMINn
TG.n
HDLCAR2n
*
−3.4
ALCOHOLNever.n
/
/
−5.09
−0.65
sin
ALBUMINNo.n
+
time.d
−
SEXfemale.n
+
tan
+
/
−1.75
sin
/
SMOKINGNever.n
sys
ALBUMINn
SMOKINGNever.n
STENOSISn
0.53
−4.35
*
−0.06
IMTn

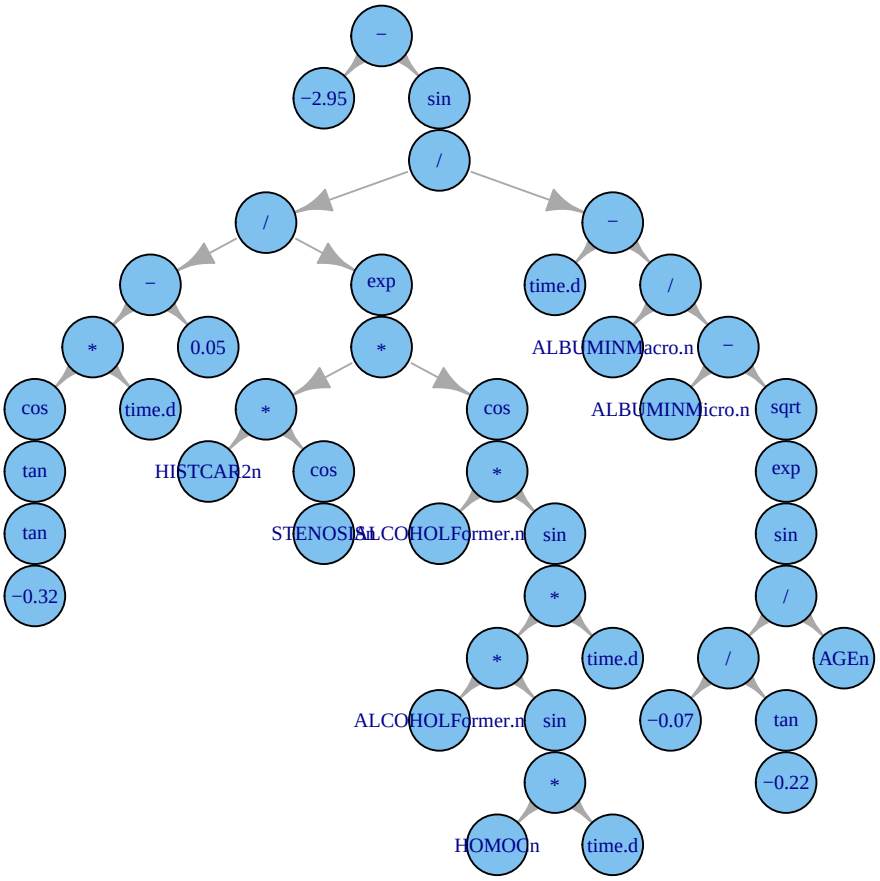

−
−2.87
cos
−
−
−
*
cos
time.d
+
*
+
/
0.06
time.d
+
tan
HISTCAR2n
−0.17
ALBUMINMacro.n
−
ALBUMINNo.n
tan
−2.87
AGEn
*
*
0.27
HOMOCn
AGEn
sin
tan
ALBUMINNo.n

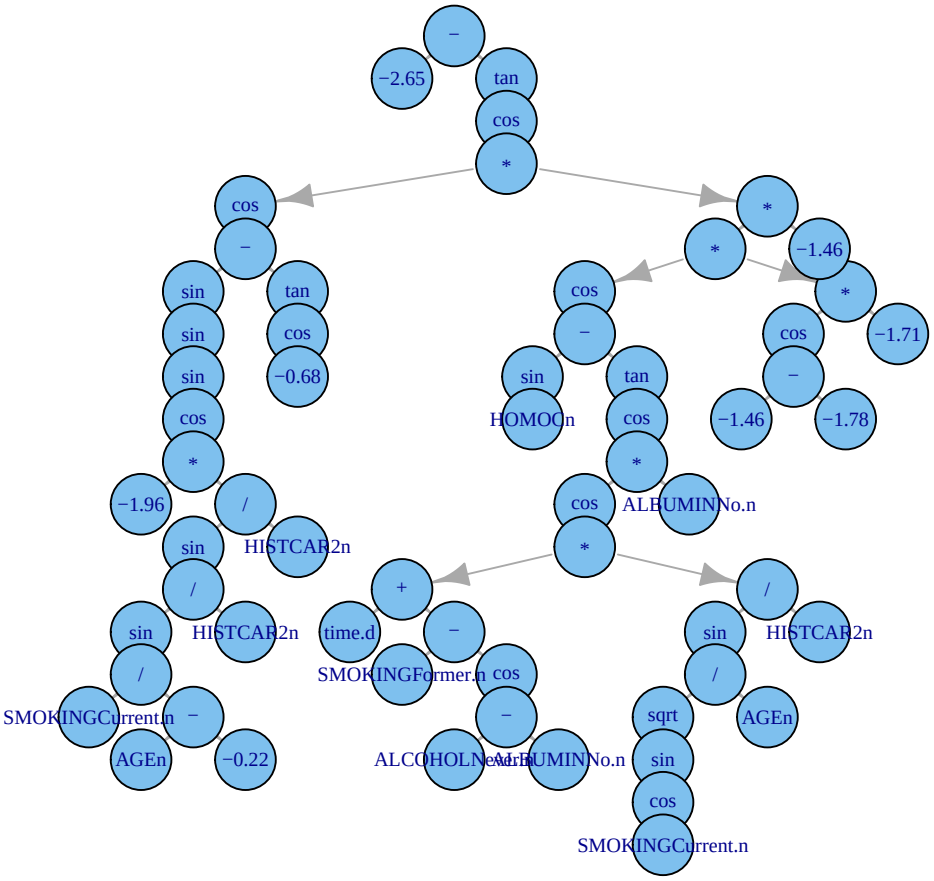

+
+
+
*
ALBUMINNo.n
sin
−3.84
ALBUMINMacro.n
*
+
HISTCAB2n
*
cos
2.26
exp
−1.08
−
*
tan
cos
ALBUMINMacro.n
*
ALBUMINNo.n
cos
1.98
*
+
exp
−1.75
time.d
sin
AGEn
+
cos
HISTCAB2n
+
SMOKINGFormer.n
HDLn

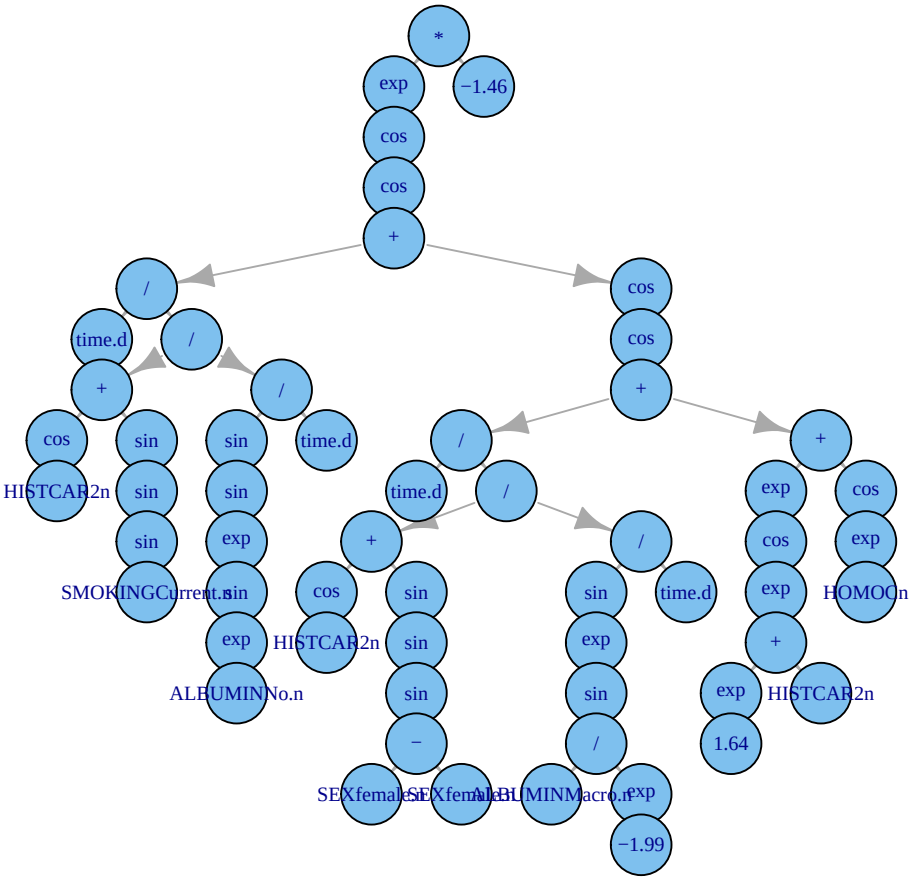

+
+
/
−2.02
cos
+
exp
+
−3.08
exp
−
cos
cos
+
CREATn
−0.46
tan
exp
+
sin
HISTCAR2n
cos
sqrt
HISTCAR2n
+
+
ALBUMINNo.n
exp
*
BMIn
STENOSISn
sin
cos
−
IMTn
time.d
*
*
/
/
/
BMIn
time.d
*
HDLn
cos
exp
time.d
AGEn
0.99
*
SYSTHn
ALBUMINMacro.n
exp
/
BMIn
sin
AGEn
sin
HISTCAR2n
IMTn

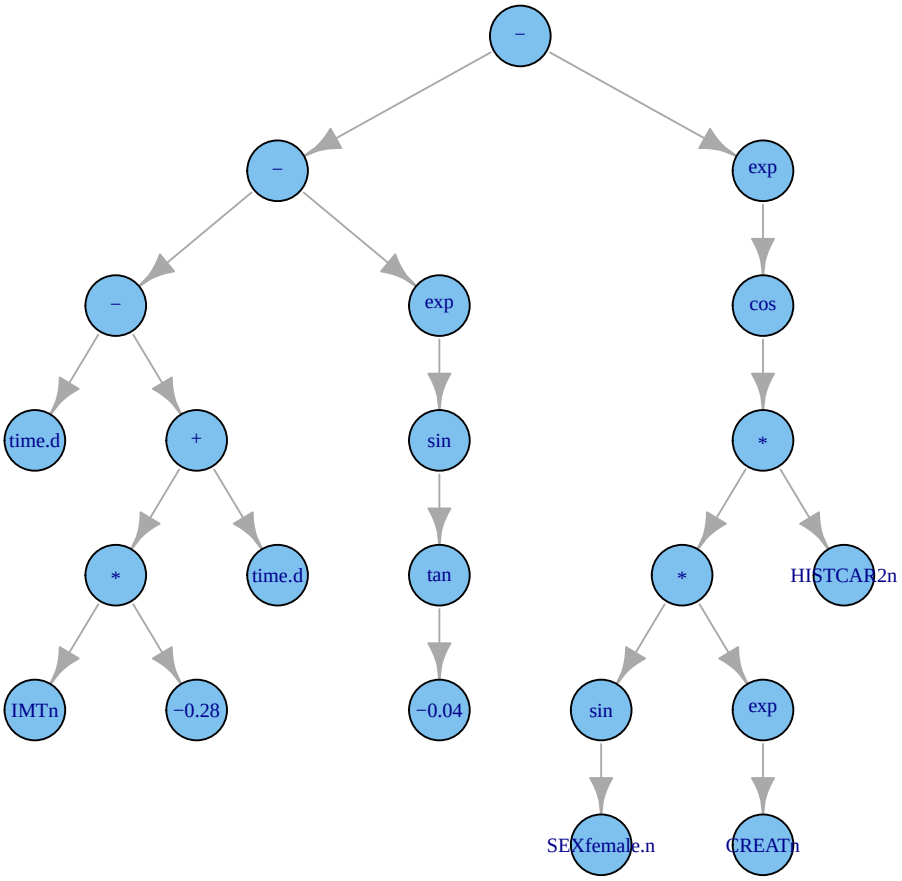

+
+
log
−2.46
0.05
*
exp
exp
2.45
*
HOMOCn
cos
exp
+
log
ALBUMINMicro.n
0.05
*
exp
exp
2.82
*
−
cos
HISTCAR2n
ALBUMINNo.n
exp
−
tan
time.d
ALBUMINNo.n
−
*
ALBUMINNo.n
*
STENOSISn
*
*
STENOSISn
HISTCAR2n
SMOKINGFormer.n
+
cos
cos
−
AGEn
*
cos
HOMOCn
−
1.24
exp
*
CREATn
SMOKINGFormer.n
SYSTHn

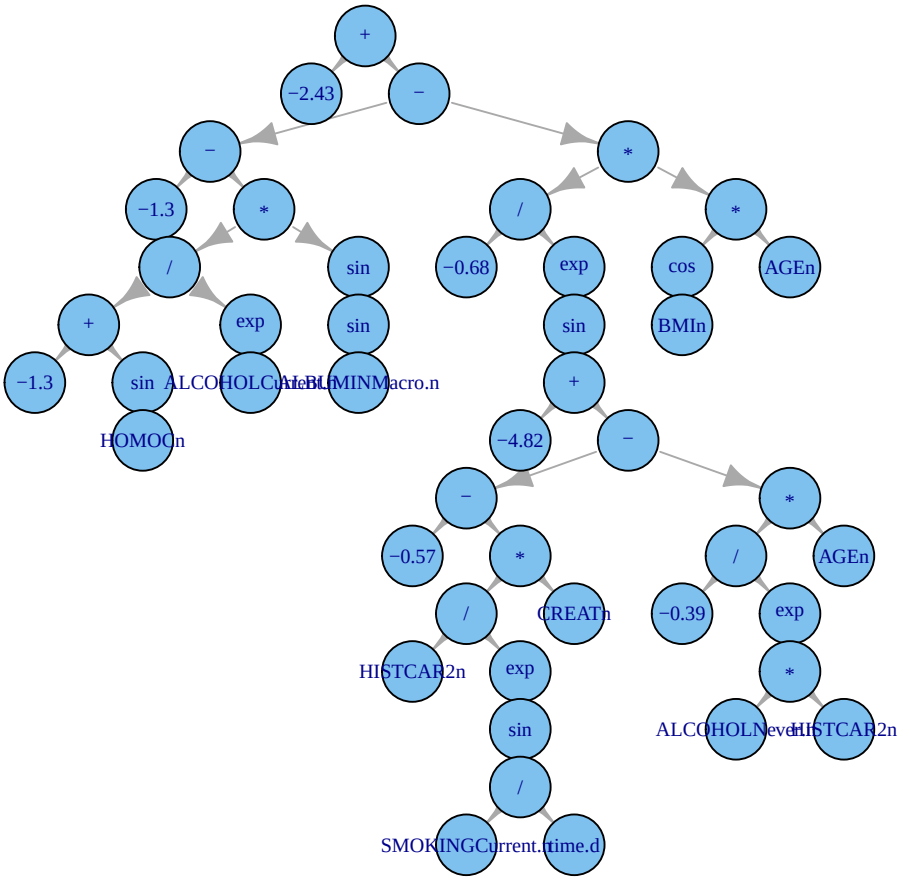

−
+
cos
sin
−2.35
*
/
ALBUMINMicro.n
/
ALBUMINMicro.n
+
AGEn
sin
cos
STENOSISn
/
exp
AGEn
sin
*
/
IMTn
*
AGEn
sin
time.d
0.09
HISTCAR2n

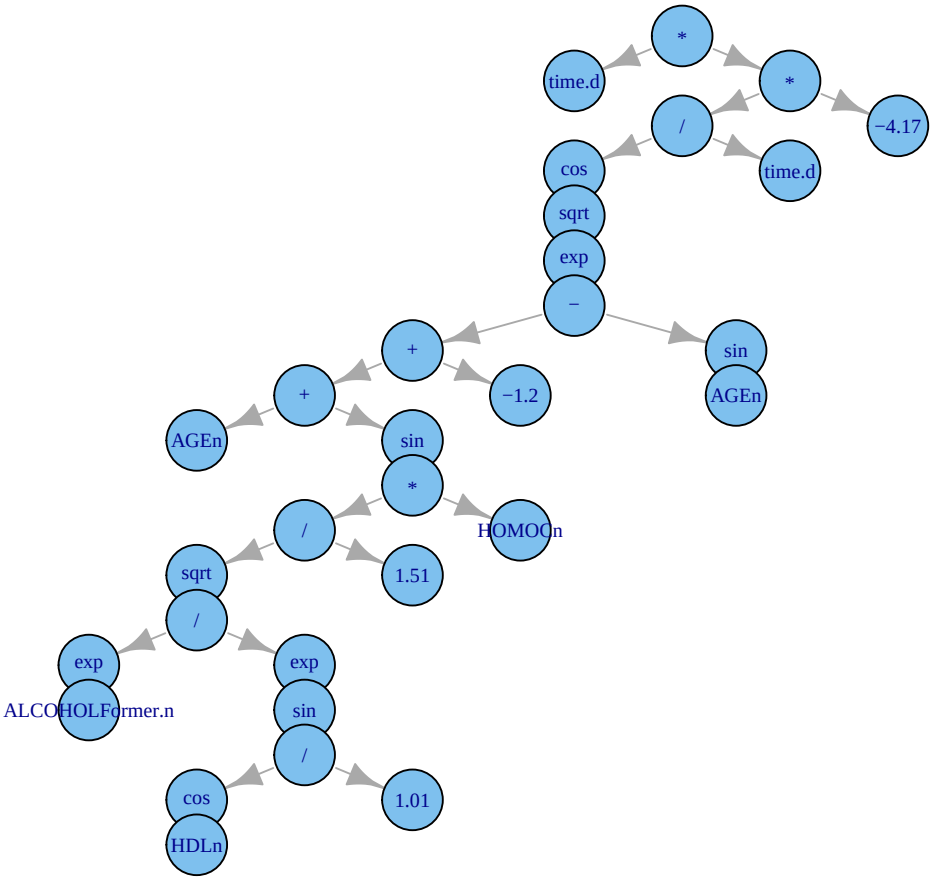

−
−2.95
sin
/
/
−
exp
−
time.d
/
*
0.05
*
ALBUMINMacro.n
−
cos
time.d
*
cos
ALBUMINMicro.n
sqrt
tan
HISTCAR2n
cos
*
exp
tan
STENOSIS1
ALCOHOLFormer.n
sin
sin
−0.32
*
/
*
time.d
/
AGEn
ALCOHOLFormer.n
sin
−0.07
tan
*
−0.22
HOMOCn
time.d

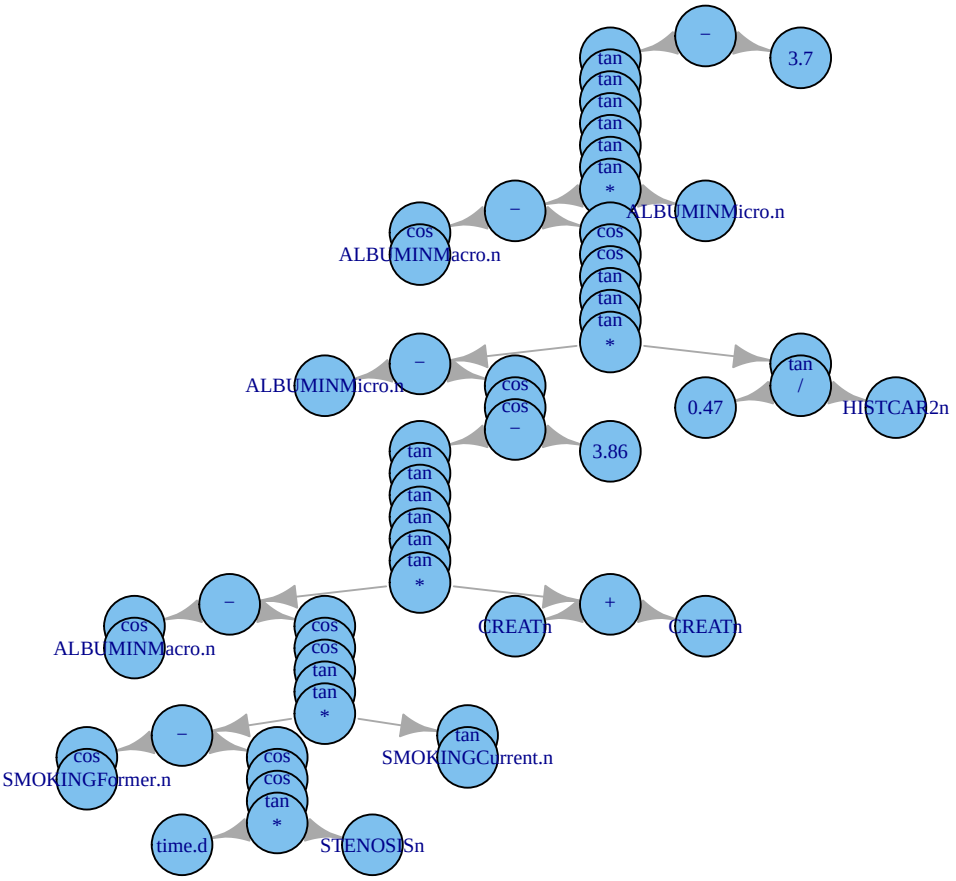

−
−2.65
tan
cos
*
cos
−
sin
sin
sin
cos
*
−1.96
/
sin
/
sin
/
SMOKINGCurrent.n
AGEn
−
−0.22
tan
cos
−0.68
HISTCAR2n
HISTCAR2n
*
cos
−
sin
HOMOCn
tan
cos
*
cos
ALBUMINNo.n
*
+
time.d
−
SMOKINGFormer.n
cos
−
ALCOHOLNeverB
ALBUMINNo.n
*
−1.46
*
cos
−
−1.46
−1.78
−1.71
/
sin
/
sqrt
sin
cos
SMOKINGCurrent.n
HISTCAR2n
AGEn

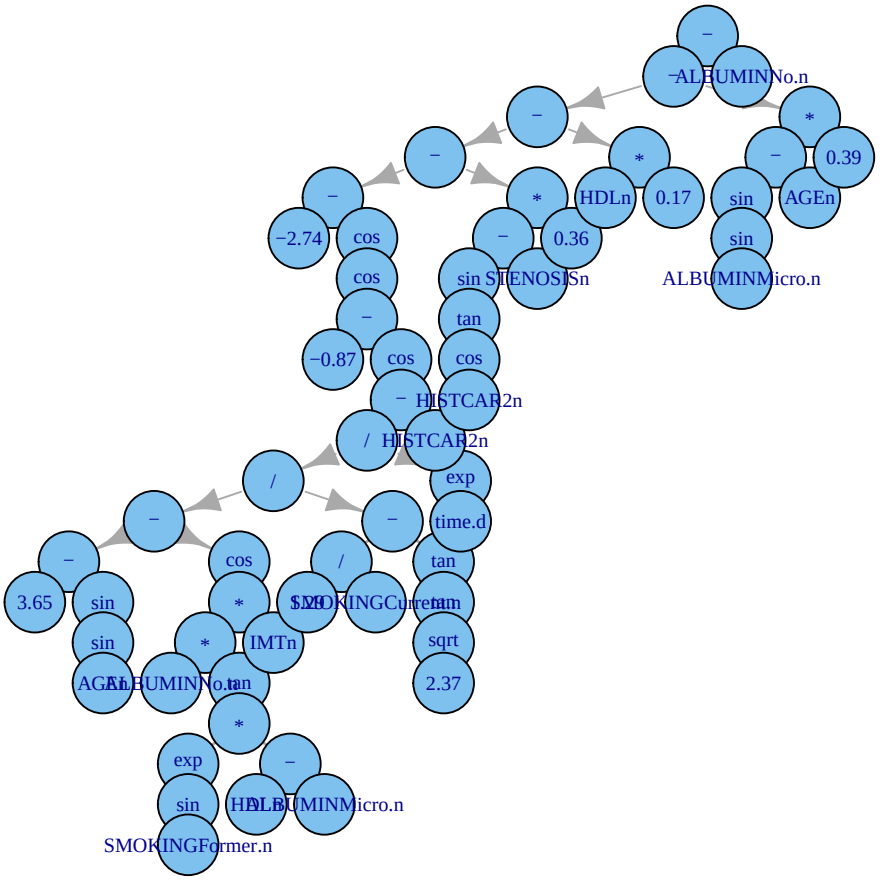

*
exp
−1.46
cos
cos
+
/
cos
time.d
/
cos
+
+
cos
sin
sin
/
/
time.d
+
HISTCAR2n
sin
sin
time.d
exp
cos
sin
sin
sin
SMOKINGCurrent.n
sin
exp
HISTCAR2n
sin
sin
sin
exp
ALBUMINNo.n
−
SEXfemale.n
SEXfemale.n
ALBUMINMacro.n
exp
−1.99
exp
cos
exp
cos
exp
HOMOCn
+
exp
HISTCAR2n
1.64

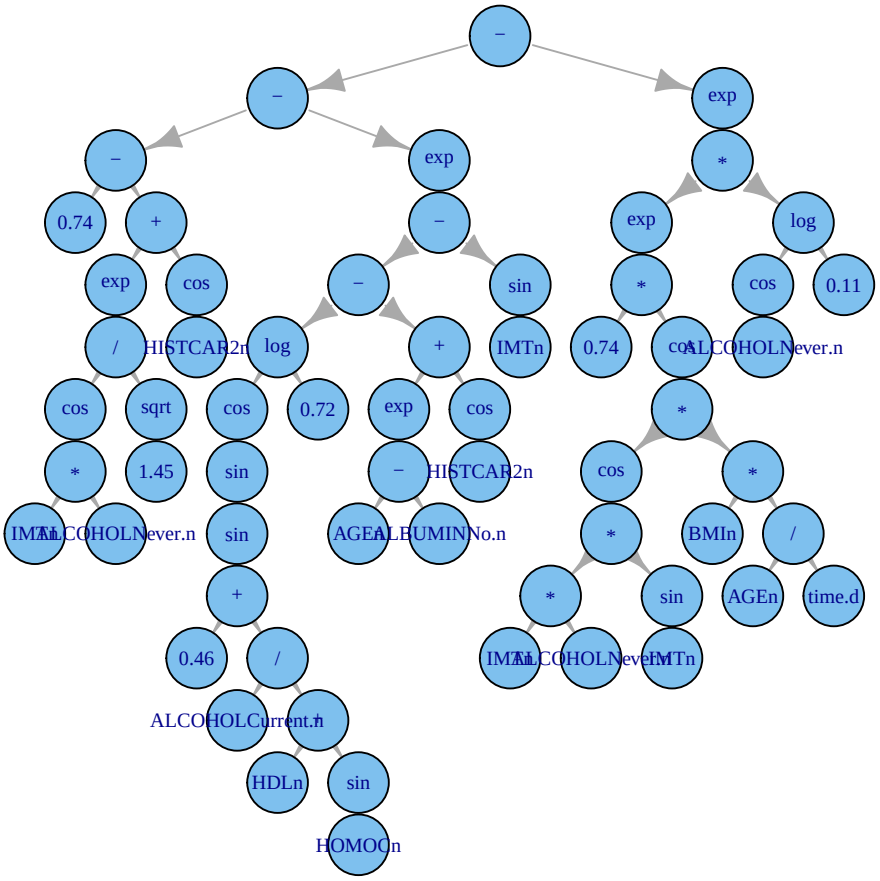

−
−
exp
−
exp
cos
time.d
+
sin
*
*
time.d
tan
*
HISTCAB2n
IMTn
−0.28
−0.04
sin
exp
SEXfemale.n
CREATn

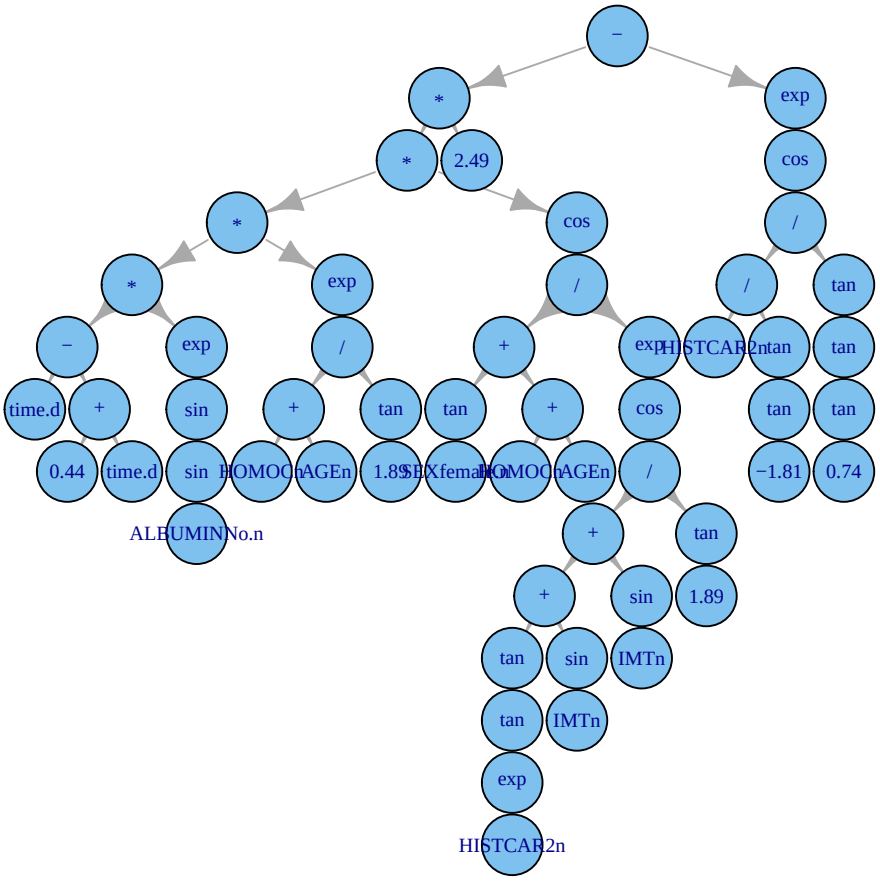

+
−2.43
−
−
−1.3
*
/
sin
+
exp
sin
−1.3
sin
HOMOCn
ALCOHOLCurrentn
ALBUMINMacro.n
*
/
*
−0.68
exp
cos
AGEn
sin
BMIn
+
−4.82
−
−
*
−0.57
*
/
AGEn
/
CREATn
−0.39
exp
HISTCAR2n
exp
*
sin
ALCOHOLNevern
HISTCAR2n
/
SMOKINGCurrent.ntime.d

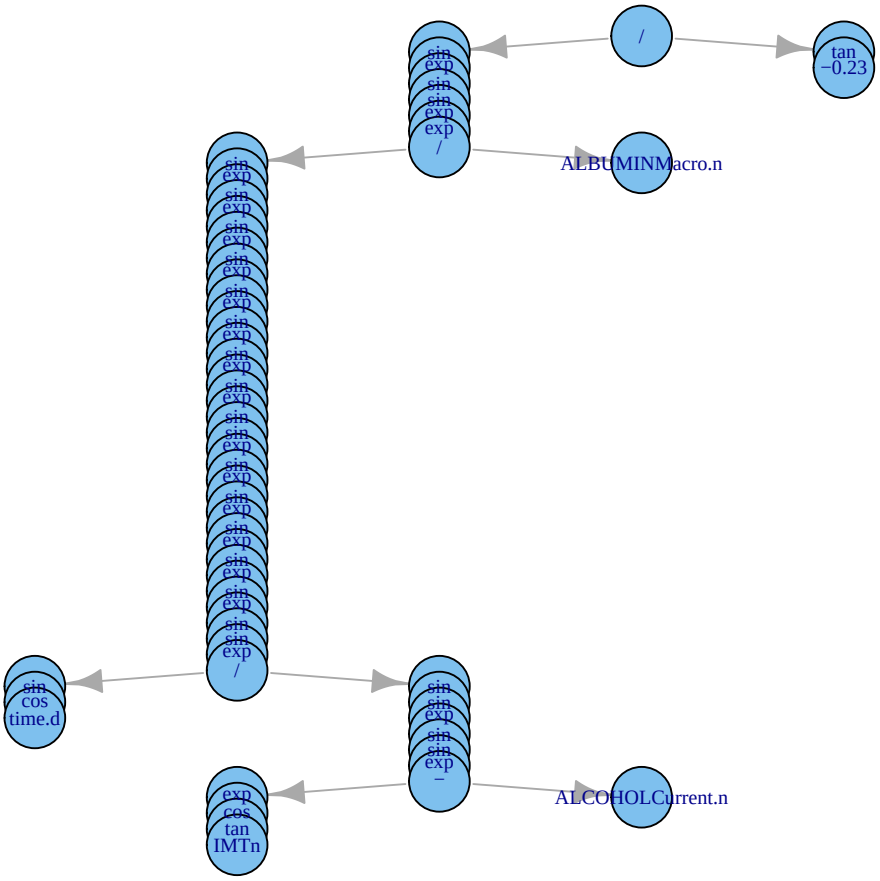

*
time.d
*
/
−4.17
cos
time.d
sqrt
exp
−
+
sin
+
−1.2
AGEn
AGEn
sin
*
/
HOMOCn
sqrt
1.51
/
exp
exp
ALCOHOLFormer.n
sin
/
cos
1.01
HDLn


−
3.7
tan
tan
tan
tan
tan
tan
*
ALBUMINMicro.n
−
cos
ALBUMINMacro.n
cos
cos
tan
tan
tan
*
−
cos
ALBUMINMicro.n
cos
−
3.86
tan
/
0.47
HISTCAB2n
tan
tan
tan
tan
tan
tan
*
−
cos
ALBUMINMacro.n
cos
cos
tan
tan
*
+
CREATn
CREATn
−
cos
SMOKINGFormer.n
cos
cos
tan
*
tan
SMOKINGCurrent.n
time.d
STENOSISn


−
−
ALBUMINNo.n
−
*
−
*
0.39
−
*
HDLn
0.17
sin
AGEn
−2.74
cos
−
0.36
sin
sin
cos
sin STENOSISn
ALBUMINMicro.n
−
tan
−0.87
cos
cos
HISTCAR2n
HISTCAR2n
/
exp
−
−
time.d
cos
/
tan
3.65
sin
*
tan
sin
*
IMTn
sqrt
AGEn
tan
2.37
*
exp
−
sin
ALBUMINMicro.n
SMOKINGFormer.n


−
−
exp
−
exp
0.74
+
−
sin
*
log
exp
cos
exp
cos
*
cos
0.11
/
HISTCAR2n
log
+
IMTn
0.74
cos
ALCOHOLNever.n
cos
sqrt
cos
0.72
exp
cos
*
cos
*
*
*
1.45
sin
−
HISTCAR2n
cos
*
*
IMTn
ALCOHOLNever.n
sin
AGE
ALBUMINNo.n
*
BMIn
/
+
*
sin
AGEn
time.d
0.46
/
IMTn
ALCOHOLNever.n
IMTn
ALCOHOLCurrent.n
HDLn
sin
HOMOCn


−
*
exp
*
2.49
cos
cos
/
*
*
exp
/
/
tan
−
exp
/
+
exp
tan
tan
time.d
+
sin
+
tan
tan
+
cos
tan
tan
0.44
time.d
sin
HOMOCn
AGEn
1.89
SEXfemaleCn
HOMOCn
AGEn
/
−1.81
0.74
ALBUMINNo.n
+
tan
+
sin
1.89
tan
sin
IMTn
tan
IMTn
exp
HISTCAR2n


/
tan
−0.23
exp
/
ALBUMINMacro.n
exp
/
sin
cos
time.d
exp
−
ALCOHOLCurrent.n
exp
cos
tan
IMTn

Appendix D

Predictor Effects: SMART experiments

Plots of the effects of predictor values on log hazard in the 'final' GP model in the SMART experiments in chapter 6

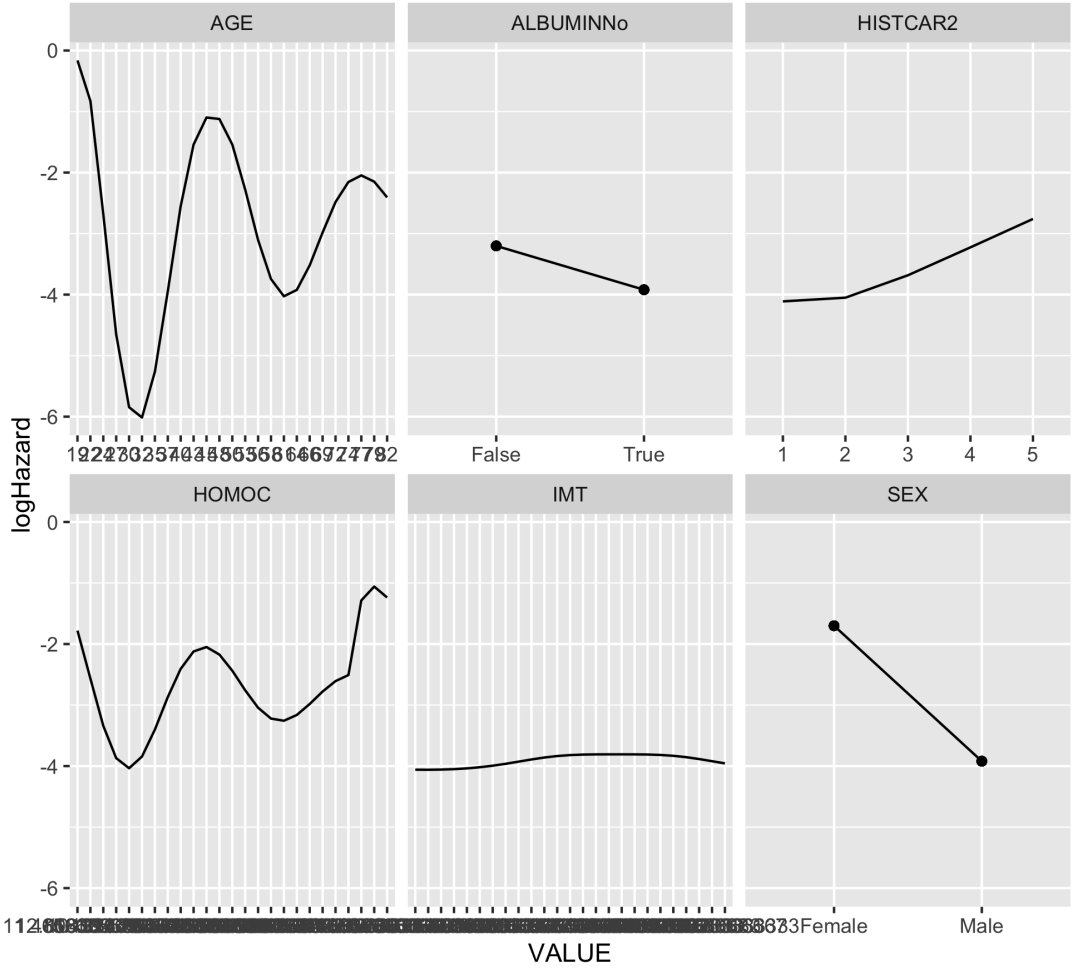


Figure D.1: Plots of the effects of predictor values on log hazard in the 'final' GP model in the SMART experiments in chapter 6.

Appendix E

Results: SMART experiments (secondary analysis)

Results of the additional experiments that repeat the SMART experiments in chapter 6, but only on the subset of covariates that were selected with a relatively high frequency (> 0.5) in the main experiment.

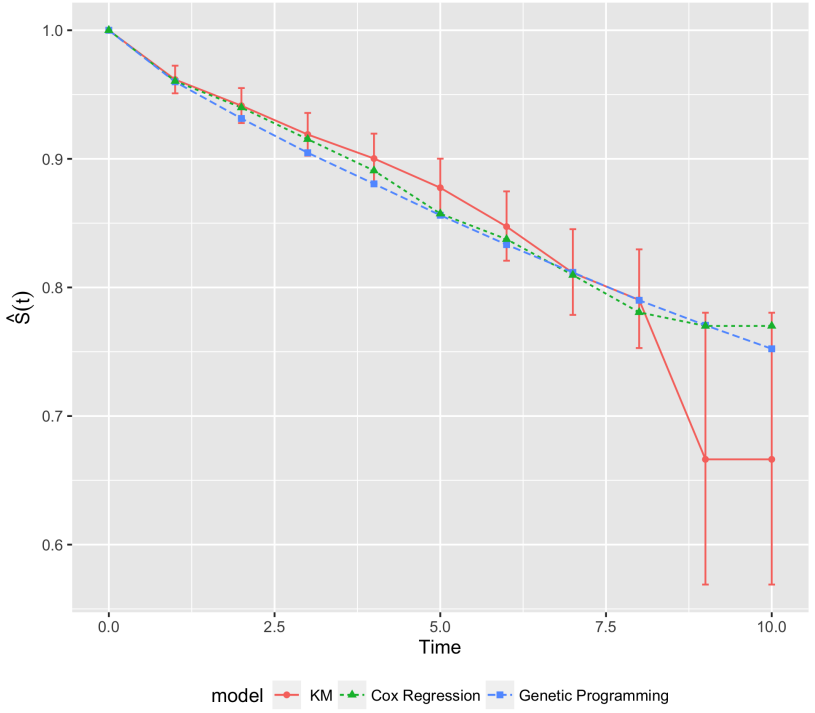


Figure E.1: Average survival curves for the Cox regression and genetic programming models. The error bars represent ± 2 standard errors of the KM estimates.

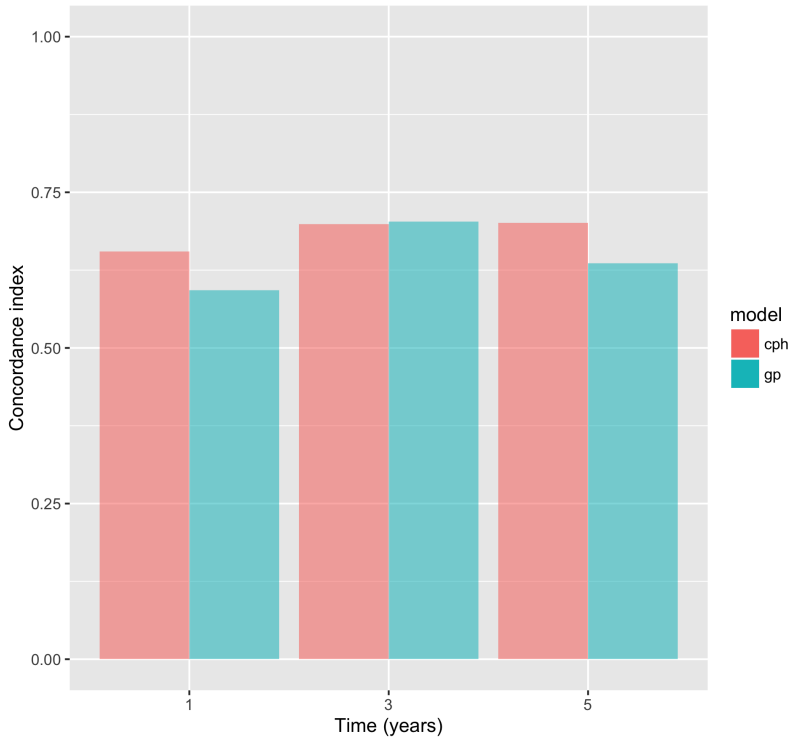


Figure E.2: C-statistic estimates by model for t=1, 3 and 5 years

Table E.1: C-statistic estimates by model at t=1, 3, and 5 years

Time (years)	Genetic Programming (superset)	Genetic Programming (subset)
1	0.59	0.59
3	0.69	0.70
5	0.64	0.64

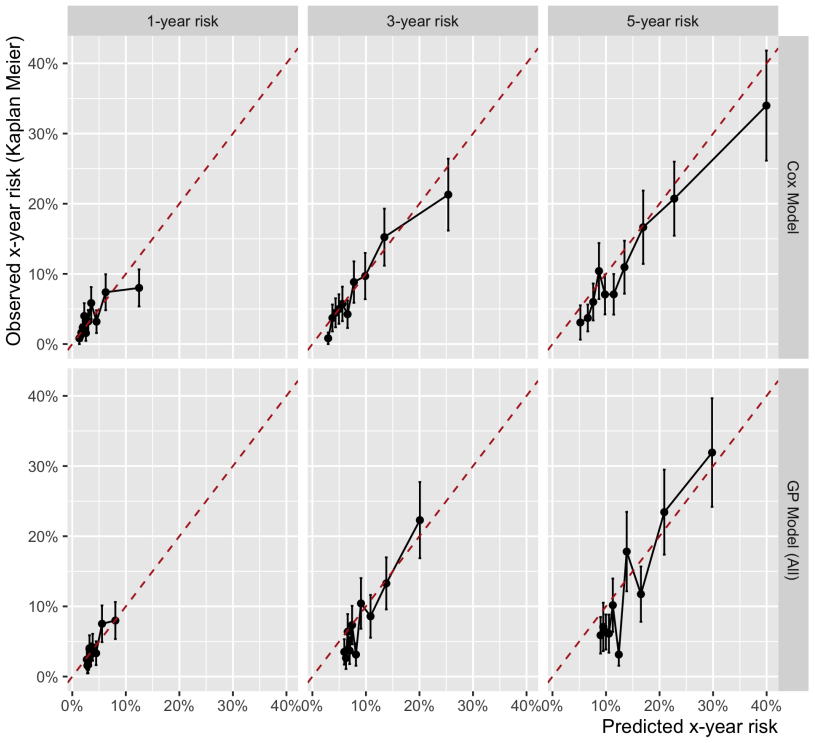


Figure E.3: Calibration plots for the Cox regression and genetic programming models, at t=1, 3, and 5 years..

Table E.2: χ^2 statistic for the comparison between observed versus expected (according to the model) number of events in groups of patients defined according to the predicted $1 - S(t)$ at t=1, 3, and 5 years.

Time (years)	Genetic Programming (superset)		Genetic Programming (subset)	
t	χ^2	p-value	χ^2	p-value
1	5.18	0.818	3.44	0.944
3	9.99	0.352	12.13	0.206
5	16.17	0.063	21.88	0.009

Appendix F

ISAC Protocol: CPRD Experiments

ISAC APPLICATION FORM
OTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

ly: mber tted	....	**IMPORTANT** **If you have any queries, please contact ISAC Secretariat:** ISAC@cprd.com

itle
oping a new cardiovascular risk score for the UK general population

l Investigator (full name, job title, organisation & e-mail address for correspondence regarding this l)
n Bannister, MSc, Postgraduate Researcher, Cardiff University, bannisterca@cf.ac.uk

on (full address)
ersity School of Medicine, Primary Care & Public Health, Cardiff Medicentre, Heath Park, Cardiff, CF14 4UJ

l's Author (if different from the principal investigator)

f Institution (please tick one box below)

hia ☒ Research Service Provider ☐ Pharmaceutical Industry ☐
☐ Government Departments ☐ Others ☐

al Sponsor of study

ceutical Industry *(please specify)* ☐ Academia *(please specify)* ☒ Cardiff ity
ment / NHS *(please specify)* ☐ None ☐
(please specify) ☐

ource *(please tick one box below)*

r has on-line access ☒ Purchase of ad hoc dataset ☐
ssioned study ☐
☐ *(please specify)*

s protocol been peer reviewed by another Committee?

☐ No ☒

ate in your protocol the name of the reviewing Committee(s) and provide an outline of the review process e.

f Study *(please tick all the relevant boxes which apply)*

g Reaction/Drug Safety ☐ Drug Use ☐ Disease Epidemiology ☒
veness ☐ Pharmacoeconomic ☐ Other ☐

dy is intended for:

tion in peer reviewed journals ☒ Presentation at scientific conference ☐
tation at company/institutional meetings ☐ Other Form
r(s) of PhD thesis

11. Does this protocol also seek access to data held under the CPRD Data Linkage Scheme?

Yes ☒ **No** ☐

12. If you are seeking access to data held under the CPRD Data Linkage Scheme, please select the source(s) of linked data being requested.

☒ **Hospital Episode Statistics** ☐ **Cancer Registry Data*** ☐ **MINAP**
☒ **ONS Mortality Data** ☒ **Index of Multiple Deprivation/ Townsend Score**
☐ **Mother Baby Link** ☐ **Other: (please specify)**

Please note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email admin@cprd.com to discuss this requirement further.

13. If you are seeking access to data held under the CPRD Data Linkage Scheme, have you already discussed your request with a member of the Research team?

Yes ☒ **No*** ☐

Please contact the CPRD Research Team on +44 (20) 3080 6383 or email admin@cprd.com to discuss your requirements before submitting your application.

Please list below the name of the person/s at the CPRD with whom you have discussed your request.
Tarita Murray Thomas (MHRA/CPRD Enquiry Reference: OCR9669)

14. Does this protocol involve requesting any additional information from GPs?

Yes* ☐ **No** ☒

*** Please indicate what will be required:**
Completion of questionnaires by the GP• **Yes** ☐ **No** ☐
Provision of anonymised records (e.g. hospital discharge summaries) **Yes** ☐ **No** ☐
Other (please describe)

• Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.

15. Does this protocol describe a purely observational study using CPRD data (this may include the review of anonymised free text)?

Yes* ☒ **No**** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*
*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

16. Does this study involve linking to patient *identifiable* data from other sources?

Yes ☐ **No** ☒

17. Does this study require contact with patients in order for them to complete a questionnaire?

Yes ☐ **No** ☒

N.B. Any questionnaire for completion by patients must be approved by ISAC before circulation for completion.

18. Does this study require contact with patients in order to collect a sample?

Yes* ☐ **No** ☒

**** Please state what will be collected***

19. Experience/expertise available

Please complete the following questions to indicate the experience/expertise available within the team of researchers actively involved in the proposed research, including analysis of data and interpretation of results

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	☐	☐
1-3	☐	☐
> 3	☒	☒

	Yes	No
Is statistical expertise available within the research team?	☒	☐

If yes, please outline level of experience *The research team has statistical expertise appropriate to use of large datasets for epidemiological research. Additional expertise can be sought from colleagues within Cardiff School of Medicine as required.*

	Yes	No
Is experience of handling large data sets (>1 million records) available within the research team?	☒	☐

If yes, please outline level of experience *The research team have extensive experience of using large routine NHS datasets including HES, GRPD and THIN.*

	Yes	No
Is UK primary care experience available within the research team?	☒	☐

If yes, please outline level of experience *The team has experience of using primary care data form a variety of studies. Specific clinical assistance can be sought if required from colleagues from the department of Primary Care and Public Health within Cardiff School of Medicine*

20. References relating to your study

Please list up to 3 references (most relevant) relating to your proposed study.

1 Anderson KM, Odell PM, Wilson PWF, KannelWB. Cardiovascular disease risk profiles.Am Heart J 1991;121(1 pt 2):293-8.

2 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J,Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England andWales: prospective derivation and validation of QRISK2. BMJ 2008;336:1475-82, doi:10.1136/bmj.39609.449676.25.

3 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J,MayM, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007;335:136, doi:10.1136/bmj.39261.471806.55.

21. List of all investigators/collaborators (*please list the names, affiliations and e-mail addresses* of all collaborators*, other than the principal investigator)

Sara Jenkins-Jones, MSc
Postgraduate Researcher, Cardiff University
s.jenkins-jones@cs.cardiff.ac.uk

Professor Craig Currie
Professor of Applied Pharmacoepidemiology, Cardiff University
currie@cardiff.ac.uk

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

PROTOCOL CONTENT CHECKLIST

In order to help ensure that protocols submitted for review contain adequate information for protocol evaluation, ISAC have produced instructions on the content of protocols for research using CPRD data. These instructions are available on the CPRD website (www.cprd.com/ISAC). All protocols using CPRD data which are submitted for review by ISAC must contain information on the areas detailed in the instructions. IF you do not feel that a specific area required by ISAC is relevant for your protocol, you will need to justify this decision to ISAC.

Applicants must complete the checklist below to confirm that the protocol being submitted includes all the areas required by ISAC, or to provide justification where a required area is not considered to be relevant for a specific protocol. Protocols will not be circulated to ISAC for review until the checklist has been completed by the applicant.

Please note, your protocol will be returned to you if you do not complete this checklist, or if you answer 'no' and fail to include justification for the omission of any required area.

	Included in protocol?		
Required area	Yes	No	If no, reason for omission
Lay Summary (max.200 words)	☒	☐	
Background	☒	☐	
Objective, specific aims and rationale	☒	☐	
Study Type *Descriptive* *Hypothesis Generating* *Hypothesis Testing*	 ☐ ☐ ☒	 ☒ ☒ ☐	Study is primarily hypothesis testing
Study Design	☒	☐	
Sample size/power calculation (Please provide justification of sample size in the protocol)	☒	☐	
Study population (including estimate of expected number of relevant patients in the CPRD)	☒	☐	
Selection of comparison group(s) or controls	☒	☐	
Exposures, outcomes and covariates *Exposures are clearly described* *Outcomes are clearly described*	☒ ☒ ☒	☐ ☐ ☐	
Data/ Statistical Analysis Plan *There is plan for addressing confounding* *There is a plan for addressing missing data*	☒ ☒ ☒	☐ ☐ ☐	
Patient/ user group involvement †	☐	☒	n/a
Limitations of the study design, data sources and analytic methods	☒	☐	
Plans for disseminating and communicating study results	☒	☐	

† ***It is expected that many studies will benefit from the involvement of patient or user groups in their planning and refinement, and/or in the interpretation of the results and plans for further work. This is particularly, but not exclusively true of studies with interests in the impact on quality of life. Please indicate whether or not you intend to engage patients in any of the ways mentioned above.***

ISAC strongly recommends that researchers using CPRD consider registering as a NRR data provider in order that others engaged in research within the UK can be made aware of current works. The ***National Research Register (NRR)*** *is a register of ongoing and recently completed research projects funded by, or of interest to, the United Kingdom's National Health Service. Information on the NRR is available on www.nrr.nhs.uk .*
Please Note: Registration with the NRR is entirely voluntary and will not replace information on ISAC approved protocols that are published in summary minutes or in the ISAC annual report*.*

GPRD ISAC protocol

Developing a new cardiovascular risk score for the UK general population

Christian Bannister, MSc
Postgraduate Researcher, Cardiff University

Sara Jenkins-Jones, MSc
Postgraduate Researcher, Cardiff University

Professor Craig Currie
Professor of Applied Pharmacoepidemiology, Cardiff University

Submitted on behalf of Cardiff University

Table of Contents

Lay summary

Cardiovascular disease (CVD) is the leading cause of mortality and a major cause of morbidity globally and in the UK. National policies support targeting of interventions to reduce risk of cardiovascular disease among high-risk patients. General practitioners use validated risk prediction models to identify these high risk patients. Numerous multivariate risk models have been developed to estimate a patient's risk of cardiovascular disease based on key known risk factors.

Until recently in the UK, the National Institute for Clinical Excellence (NICE) recommended the long-established Framingham equation for cardiovascular risk prediction. Published reviews have suggested that Framingham may over-estimate risk by up to 50% in contemporary northern European populations. NICE have now ceased recommendation of any single model, leaving practitioners to decide which model to use. The purposes of the proposed study is develop a new model to predict cardiovascular risk that is more closely calibrated to the UK general population and to validate its performance against current risk models used in UK general practice.

Introduction

Cardiovascular disease (CVD) is a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, pulmonary embolism, hypertension and heart failure. The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol[1]. CVD is the leading cause of death globally: [1]. An estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to cerebrovascular disease. By 2030, almost 23.6 million people will die from CVD, mainly from heart disease and cerebrovascular disease. These are projected to remain the single leading causes of death[1].

Asymptomatic patients that are suspected to be at high risk need to be identified by General Practitioners so they can offer advice about lifestyle changes and initiate preventative treatment. To facilitate this, General Practitioners need tools that can accurately and reliably predict cardiovascular risk in their patients. National policies now support targeting of interventions to reduce risk of cardiovascular disease among high-risk patients[2-5]. There are numerous risk models that have been developed to predict the risk of cardiovascular outcomes for a 10-year time horizon based using key known risk factors. Such models include the Framingham Risk Score[6] and the Reynolds Risk Score[7], both developed from US data. The SCORE risk function was developed using data from various European countries[8]; ASSIGN originated from Scottish data[9]; while QRISK2[10-12] and the Joint British Society (JBS)[13] models were computed from UK data. Framingham is the most commonly used model in the UK[6] but has some well-documented limitations; namely that it is poorly calibrated to an ethnically diverse population such as the UK[14]. Framingham may perform well in ethno-demographically similar populations to its source but may over-estimate risk by up to 50% in contemporary northern European populations[20].

Thus far the models discussed have all been examples of validated 10-year risk models with an absolute risk threshold of 20% specified by the NICE[3]. Patients with a 10-year score above 20% are considered high risk and targeted for primary prevention measures. Applying this 20% risk threshold for intervention may not identify younger patients who, because of their age, have a low absolute 10-year risk but who have a high relative risk compared with their peers[16]. This is because age has such a dominant effect in calculating absolute cardiovascular risk. Some argue that younger patients with an adverse risk profile may have more to gain during their lifetime if interventions are started at a younger age rather than waiting until they cross the 20% threshold[17-20]. However lifetime risk models predict the cumulative risk of the event of interest over the remainder of the patient's life[21] and may provide a more appropriate assessment of future risks, particularly for younger ages[16-19].

There is currently only a single published model[16] that predicts lifetime risk of cardiovascular disease derived from contemporary UK data, and no mention of such models in UK guidelines.

A recent systematic review of cardiovascular risk models[37] reported that in none of the 21 risk scores reviewed was the effect of treatment fully assessed or adjusted for. The review suggests that two treatment effects need to be considered: (1) prior treatment (started before enrolment in the study) and (2) subsequent treatment started during study follow-up (treatment drop-ins). None of the risk scores addressed the effect of subsequent treatment[37].

The current body of research and a lack of consensus on which risk score is most appropriate for UK general practice suggests the need for cardiovascular risk models that are closely calibrated to the contemporary UK population and can adequately account for treatment effects. Such models would enable GPs to not only understand the risk of a certain outcome but also understand the changes in risk resulting from changes in treatment. Models that could adequately account for treatment effects would be a first, making a contribution to this body of research and avoid some of limitations of the existing models. The proposed study requests approvals to use Clinical Practice Research Datalink (CPRD) data to derive and validate new cardiovascular risk scores that, whilst accounting for treatment effects, provide accurate estimates of cardiovascular risk in patients from different ethnic groups in the UK general population.

Objective

The objective of the proposed study is to develop and validate new cardiovascular risk algorithms that, whilst accounting for treatment effects, provide accurate estimates of cardiovascular risk in patients from different ethnic groups in the UK general population.

Methods

Data source
CPRD with linked HES, ONS and Census datasets.

Study type
The study will be primarily hypothesis testing with regard to cardiovascular risk in the general population.

Study design
The study will use a prospective open cohort design.

Study population and cohorts
The study proposes a single open cohort from the general population, over 14 years from 1997 to 2011. To ensure completeness of recording of morbidity and prescribing data, practices in CPRD that have been 'Up-To-Standard' (UTS) for at least one year will be considered eligible for inclusion. An open cohort of all patients, aged 35-74 years at index date, drawn from all patients registered with eligible practices from 1 January 1997 to 31 December 2011. Index date will be the latest of the following dates: mid-year estimate of 35th birthday, date of registration with the practice, practice UTS date, and the beginning of the study period (1 January 1997). In addition we will only include patients in the analysis once they have a minimum of one year's complete data after the UTS date.

Patients will be excluded from the cohort that have any one of the following criteria: a recorded diagnosis of cardiovascular or cerebrovascular disease prior to the index date; any temporary residence status; interrupted periods of registration with the practice; no valid Townsend Score; were taking statins at index; implausible or improbable dates; or recorded risk factor values out of plausible range. A 'wash-in' period of 365 days will be applied prior to the index to further ensure excluding prior history of cardiovascular or cerebrovascular disease and that patients weren't taking statins at baseline. Patients will be selected that are eligible for linkage schemes with the Hospital Episode Statistics (HES), Office of National Statistics (ONS) mortality data and Index of Multiple

Deprivation/ Townsend Score data throughout their respective period of follow-up. This should provide accurate ascertainment of ethnicity, socioeconomic status, and cause of death. Issues with ethnicity data where within the non-missing data there are a large proportion of ethnicities recorded as 'unknown' will be addressed by recoding the 'unknown' responses as 'white', with the rationale that, assuming the study population is comparable with the UK population, 93% or more of people without ethnicity recorded would be expected to be from a white ethnic group. Linkage with MINAP was considered but its merits were thought to be outweighed by the reduction in cohort size and study period from linkage scheme eligibility constraints. Table 1 below provides an estimate of expected number of eligible patients for this study.

Stage	Description	Lost	Remaining
1	CPRD patients with status 'accept' (CPRD quality indicator)	n/a	11,801,879
2	Omit patients in 1 where ineligible for required linkage schemes	9,537,894	2,263,985
3	Omit patients in 2 where age at eligibility for entry into study between 35 and 74	133,242	2,130,743
4	Omit patients in 3 where date at eligibility for entry between '1997-01-01' and '2011-12-31'	1,165,441	965,302
5	Omit patients in 4 taking statins at baseline	287,641	677,661
6	Omit patients in 5 with CVD at baseline	18,143	659,518
7	Omit patients using a 'wash-in' of 365 days for 5 and 6	24,669	634,849

Table 1: Attrition from cohort selection criteria indicating expected numbers of eligible patients

Sample size and power calculation
A sample size calculation has not been carried *a priori* as this the study does not intend to take a sample *per se* but rather to use all available data in GPRD that meets the selection criteria.

Selection of comparison group(s) or controls
A split-sample (or cross-validation) approach will be taken for model development, 70% of the eligible patients shall be randomly allocated to a training set and 30% will be randomly allocated to a test set. Based on estimates of study cohort size from table 1, expected numbers of patients in the training and test sets would be 444,394 and 190,455 respectively. Data from the training set will be used to derive the new risk score and the test used to validate it. Although this method provides some assurance that the model will 'overfit' the data, this approach is not an adequate surrogate for external validation.

Outcomes
Primary outcome measure is first recorded diagnosis of cardiovascular disease recorded by the general practice either before or at death or via their linked ONS death certificate within the study period. Cardiovascular disease is defined as coronary heart disease (myocardial infarction, angina), stroke or transient ischaemic attacks in the term cardiovascular disease but not peripheral vascular disease. Read code definitions will be used for case identification of coronary heart disease and cerebrovascular disease in general practice records. ICD-9 and ICD-10 codes will be used for case identification in HES data and on ONS death certificates. For HES we will consider patients admitted as emergencies with a primary diagnosis of the relevant event. For ONS we will consider deaths with a primary or contributory cause of death of coronary heart disease or cerebrovascular disease. A preliminary set of outcome codes for this study have been included as appendix A.

Model covariates

Covariates proposed for model derivation include the following (where appropriate these risk factors will be modelled as time-independent, time-dependent or both):

1. Age (years)
2. Self assigned ethnicity (white/not recorded, Indian, Pakistani, Bangladeshi, other Asian, black African, black Caribbean, Chinese, other including mixed)
3. Sex (male v female)
4. Smoking status (current smoker, non-smoker (including ex-smoker))
5. SBP (continuous)
6. Lipid Profile:
    a. Total serum cholesterol (continuous)
    b. High-density lipids (continuous)
    c. Low-density lipids (continuous)
    d. Triglycerides (continuous)
    e. Ratio of total serum cholesterol: HDL cholesterol (continuous)
7. BMI (continuous)
8. Family history of CHD in first-degree relative under 60 years (yes/no)
9. Townsend deprivation score (output area level 2001 census data evaluated as a continuous variable)
10. Treated hypertension (diagnosis of hypertension and at least one current prescription of at least one antihypertensive agent-e.g. thiazide, β blocker, calcium channel blocker, or angiotensin converting enzyme inhibitor) (yes/no)
11. T2DM (yes/no)
12. Renal disease (yes/no)
13. Atrial fibrillation (yes/no)
14. Rheumatoid arthritis (yes/no)
15. Charleston Index (categorical)
16. No. GP attendances in year prior
17. Angiotensin-converting-enzyme inhibitor or angiotensin receptor blockers (yes/no)
18. β -blockers (yes/no)
19. Anti-Platelet Therapy (yes/no)
20. Statin therapy (categorical)
21. Other Lipid-lowering therapies (yes/no)
22. Other cardiovascular disease (yes/no)

Prior treatment effects will be accounted for by time-independent covariates capturing status at baseline. Where appropriate, subsequent treatment effects will be accounted for by time-dependent covariates in two forms: ever exposed (e.g. was the patient exposed to therapy x at any time during the study period) and cumulative effect (e.g. over how many time periods, if any, was the patient exposed to therapy x during the study period).

Data/statistical analysis

Descriptive data on the study population shall be evaluated to determine the comparability of the randomly assigned training and test groups and evaluate the likelihood of any selection bias or confounding. Baseline characteristics will also be evaluated to help determine the generalisability of the study population to other populations. Declining secular trends in cardiovascular mortality and all-cause mortality in the general population shall be addressed by performing a sensitivity analysis using 5-year periods. Where appropriate missing data shall be handled using multiple imputation. Multiple imputation is a powerful technique that offers substantial improvements over the biased and flawed value replacement approaches based on complete cases or cases matched for age and sex[22-23]. It involves creating multiple copies of the data and imputing the missing values with sensible values

randomly selected from their predicted distribution. We propose to use the Multivariate Imputation by Chained Equations (MICE) approach.

All the risk scores mentioned thus far[6-13] have been developed using regression techniques, namely survival analysis. The study proposes two regression approaches for the derivation of a new cardiovascular risk model, the long-established survival model methods and a newer regression method called symbolic regression.

Survival Analysis
The study proposes the use of survival models (such as Cox proportional hazards or parametric survival models) on the training dataset to estimate the coefficients associated with each potential risk factor for the prediction of the outcome of interest (first recorded diagnosis of cardiovascular disease) for men and women separately. We have proposed a set of potential variables *a priori* but will consider these variables for inclusion in the model using significance tests and models will be compared using likelihood measures such as Akaike Information Criterion (AIC). The proportional hazards assumptions will be checked for each variable and tested for any non-linear relation between continuous independent variables and the outcome. Covariate interactions effects will also be considered for inclusion in the model. If there are missing values for key predictors then multiple imputation will be performed.

The coefficients (i.e. weights for the new cardiovascular disease risk equation) from the final survival model, built using the training dataset, will be used to obtain the predicted risk scores for all patients in the previously unseen test dataset. These predicted risk scores from the test dataset shall then be used to evaluate its performance of the final model in terms of calibration and discrimination. Calibration refers to how closely the predicted risk of cardiovascular disease agrees with the observed risk for a certain period. This will be assessed for each tenth of predicted risk, ensuring 10 equally sized groups, and for each 5 year age band by calculating the ratio of predicted to observed risk of cardiovascular disease separately for men and for women. Calibration of the risk score predictions will be assessed by plotting observed proportions versus predicted probabilities. Calibration measures such as Brier Score (adjusted version for censored data)[24] and R^2 statistic[26] will also be calculated. Discrimination is the ability of the risk score to differentiate between patients who experience a cardiovascular event during the study and those who do not. This shall be measured using the D statistic[25] and Area Under the Receiver Operating Characteristic (ROC) Curve statistic[27]. As in the training set, missing values in the test dataset will be imputed as required.

Symbolic Regression
Genetic programming (GP) is an evolutionary computation technique that automatically solves problems without requiring the user to know or specify the form or structure of the solution in advance. At the most abstract level GP is a systematic, domain-independent method that allows computers to solve problems automatically starting from a high-level statement of the problem [28].

Symbolic regression attempts to find a function that fits the given data points without making any assumptions about the structure of that function. Since GP makes no such assumption, it is well suited to this sort of discovery task[28]. Symbolic regression was one of the earliest applications of GP[29], and continues to be widely studied[30-33]. Although computationally intensive, GP is well suited to large datasets such as CPRD with the ability to handle a large number of variables and/or cases, inherently performing feature selection (i.e. selection of which variables should be included in the models) leading to parsimonious models. The data and analysis steps that are proposed for the survival models will be the same for the symbolic regression. The only difference will be that the model itself, the risk equation, will be developed using the symbolic method.

Model Comparisons

We propose to compare the performance of the derived survival and symbolic regression models on the test dataset. We shall calculate the predicted cardiovascular risk for each patient in the test dataset using the developed survival and symbolic regression models. We shall then calculate the mean predicted risk for each model and the observed risk from the data. We shall then compare the predicted and observed risk by 10^{th} of predicted risk for each score. Observed risk will be obtained using Kaplan-Meier estimates. We also propose to use the 20% risk threshold as specified by NICE guidelines [3] to calculate and compare the proportions of patients that would be classified as at 'high-risk' by each model.

Limitations of the study design, data sources and analytic methods

This study has a number of limitations. CPRD collates data from routine practice, thus there are missing and erroneous data, coding imperfections, lack of standardisation of biochemical measures (such as lipid profiles), variations between biochemical test centres and measurements are taken with varying periodicity. Certain covariates of interest such as smoking status, BMI, lipids, family history of CVD may not be recorded consistently within CPRD. There are limitations to the size of the cohort available for consideration as only a subset of CPRD patients are eligible for the data linkage schemes required. There are also limitations with ethnicity data where even within the non-missing data there are a large proportion of ethnicities recorded as 'unknown'. Limitations arise from recording the 'unknown' responses as 'white', with the rationale that, assuming the study population is comparable with the UK population, 93% or more of people without ethnicity recorded would be expected to be from a white ethnic group. Removal or exclusion of patients with missing data may introduce bias into the study; this can often be addressed by using multiple imputation techniques to impute missing values where appropriate. There are also limitations in specifying input variables *a priori* as there is potential to miss important factors and relationships that exist with variables not considered. There are limitations on the split-sample validation approach where predictive accuracy estimates, although unbiased, can be imprecise. There are also limitations inherent in all statistical modeling techniques each of which have their own set of assumptions, such as non-informative censoring, linearity, additively, proportionality, etc., that need to be satisfied in order to order to take a given approach. Violation of such statistical assumptions may preclude the use of certain techniques and/or consideration of all covariates. The proposed survival analysis approach has many such assumptions that will need to be satisfied whereas the Symbolic regression via GP has no such limitations. The computational expense such as available memory and processing speed (and thus time) may also be another limiting factor where certain techniques, main effects and n-way interaction effects cannot be considered within such computational constraints. The GP approach to symbolic regression is considered computationally expensive whereas survival analysis is computationally cheap in comparison. There may also be bias introduced due to QOF incentivisation, where general practices are paid according to the reporting under certain criteria, which will be explored using sensitivity analyses.

Plans for disseminating and communicating study results

Findings from this study will be disseminated through scientific meetings, peer-reviewed manuscript(s) and form part of a PhD thesis.

Appendix A – Preliminary Outcome Codes

code	type	endpoint	description
410	ICD9	CHD	Acute myocardial infarction
410.0	ICD9	CHD	MI, acute, anterolateral
410.1	ICD9	CHD	MI, acute, anterior, NOS
410.2	ICD9	CHD	MI, acute, inferolateral
410.3	ICD9	CHD	MI, acute, inferoposterior
410.4	ICD9	CHD	MI, acute, other inferior wall, NOS
410.5	ICD9	CHD	MI, acute, other lateral wall
410.6	ICD9	CHD	MI, acute, true posterior
410.7	ICD9	CHD	MI, acute, subendocardial
410.8	ICD9	CHD	MI, acute, spec.
410.9	ICD9	CHD	MI, acute, unspec.
411	ICD9	CHD	Other acute and subacute forms of ischemic heart disease
411.0	ICD9	CHD	Postmyocardial infarction syndrome
411.1	ICD9	CHD	Intermediate coronary syndrome
412	ICD9	CHD	Old myocardial infarction
413	ICD9	CHD	Angina pectoris
413.0	ICD9	CHD	Angina decubitus
413.1	ICD9	CHD	Prinzmetal angina
414	ICD9	CHD	Other forms of chronic ischemic heart disease
414.0	ICD9	CHD	Coronary atherosclerosis
414.1	ICD9	CHD	Aneurysm and dissection of heart
414.10	ICD9	CHD	Aneurysm of heartwall
414.11	ICD9	CHD	Aneurysm of coronary vessels
414.12	ICD9	CHD	Dissection of coronary artery
414.8	ICD9	CHD	Ischemic heart disease, chronic, other
414.9	ICD9	CHD	Ischemic heart disease, chronic, unspec.
I20	ICD10	CHD	Angina pectoris
I20.0	ICD10	CHD	Unstable angina
I20.1	ICD10	CHD	Angina pectoris with documented spasm

I20.8	ICD10	CHD	Other forms of angina pectoris
I20.9	ICD10	CHD	Angina pectoris, unspecified
I21	ICD10	CHD	Acute myocardial infarction
I21.0	ICD10	CHD	Acute transmural myocardial infarction of anterior wall
I21.1	ICD10	CHD	Acute transmural myocardial infarction of inferior wall
I21.2	ICD10	CHD	Acute transmural myocardial infarction of other sites
I21.3	ICD10	CHD	Acute transmural myocardial infarction of unspecified site
I21.4	ICD10	CHD	Acute subendocardial myocardial infarction
I21.9	ICD10	CHD	Acute myocardial infarction, unspecified
I22	ICD10	CHD	Subsequent myocardial infarction
I22.0	ICD10	CHD	Subsequent myocardial infarction of anterior wall
I22.1	ICD10	CHD	Subsequent myocardial infarction of inferior wall
I22.8	ICD10	CHD	Subsequent myocardial infarction of other sites
I22.9	ICD10	CHD	Subsequent myocardial infarction of unspecified site
I23	ICD10	CHD	Certain current complications following acute myocardial infarction
I23.0	ICD10	CHD	Haemopericardium as current complication following acute myocardial infarction
I23.1	ICD10	CHD	Atrial septal defect as current complication following acute myocardial infarction
I23.2	ICD10	CHD	Ventricular septal defect as current complication following acute myocardial infarction
I23.3	ICD10	CHD	Rupture of cardiac wall without haemopericardium as current complication following acute myocardial infarction
I23.4	ICD10	CHD	Rupture of chordae tendineae as current complication following acute myocardial infarction
I23.5	ICD10	CHD	Rupture of papillary muscle as current complication following acute myocardial infarction
I23.6	ICD10	CHD	Thrombosis of atrium, auricular appendage, and ventricle as current complications following acute myocardial infarction
I23.8	ICD10	CHD	Other current complications following acute myocardial infarction
I24	ICD10	CHD	Other acute ischaemic heart diseases
I24.0	ICD10	CHD	Coronary thrombosis not resulting in myocardial infarction
I24.1	ICD10	CHD	Dressler syndrome
I24.8	ICD10	CHD	Other forms of acute ischaemic heart disease
I25	ICD10	CHD	Chronic ischaemic heart disease
I25.0	ICD10	CHD	Atherosclerotic cardiovascular disease, so described
I25.1	ICD10	CHD	Atherosclerotic heart disease
I25.2	ICD10	CHD	Old myocardial infarction
I25.3	ICD10	CHD	Aneurysm of heart

I25.4	ICD10	CHD	Coronary artery aneurysm
I25.5	ICD10	CHD	Ischaemic cardiomyopathy
I25.6	ICD10	CHD	Silent myocardial ischaemia
I25.8	ICD10	CHD	Other forms of chronic ischaemic heart disease
I25.9	ICD10	CHD	Chronic ischaemic heart disease, unspecified

medcode	Readcode	endpoint	description
35674	14A3.00	CHD	H/O: myocardial infarct <60
40399	14A4.00	CHD	H/O: myocardial infarct >60
6336	14A5.00	CHD	H/O: angina pectoris
50372	14AH.00	CHD	H/O: Myocardial infarction in last year
57062	14AJ.00	CHD	H/O: Angina in last year
7783	323..00	CHD	ECG: myocardial infarction
39904	3232.00	CHD	ECG: old myocardial infarction
26972	3234.00	CHD	ECG:posterior/inferior infarct
55401	3235.00	CHD	ECG: subendocardial infarct
52705	3236.00	CHD	ECG: lateral infarction
39584	3889.00	CHD	Euroscore for angina
13185	662K.00	CHD	Angina control
19542	662K000	CHD	Angina control - good
15373	662K100	CHD	Angina control - poor
14782	662K200	CHD	Angina control - improving
29300	662K300	CHD	Angina control - worsening
15349	662Kz00	CHD	Angina control NOS
5904	792..00	CHD	Coronary artery operations
737	792..11	CHD	Coronary artery bypass graft operations
18249	7920.00	CHD	Saphenous vein graft replacement of coronary artery
8312	7920.11	CHD	Saphenous vein graft bypass of coronary artery
8679	7920000	CHD	Saphenous vein graft replacement of one coronary artery
7634	7920100	CHD	Saphenous vein graft replacement of two coronary arteries
7442	7920200	CHD	Saphenous vein graft replacement of three coronary arteries

11610	7920300	CHD	Saphenous vein graft replacement of four+ coronary arteries
7137	7920y00	CHD	Saphenous vein graft replacement of coronary artery OS
51515	7920z00	CHD	Saphenous vein graft replacement coronary artery NOS
9414	7921.00	CHD	Other autograft replacement of coronary artery
7134	7921.11	CHD	Other autograft bypass of coronary artery
44561	7921000	CHD	Autograft replacement of one coronary artery NEC
19413	7921100	CHD	Autograft replacement of two coronary arteries NEC
10209	7921200	CHD	Autograft replacement of three coronary arteries NEC
42708	7921300	CHD	Autograft replacement of four of more coronary arteries NEC
61310	7921y00	CHD	Other autograft replacement of coronary artery OS
7609	7921z00	CHD	Other autograft replacement of coronary artery NOS
31556	7922.00	CHD	Allograft replacement of coronary artery
32651	7922.11	CHD	Allograft bypass of coronary artery
70111	7922000	CHD	Allograft replacement of one coronary artery
57241	7922100	CHD	Allograft replacement of two coronary arteries
45886	7922200	CHD	Allograft replacement of three coronary arteries
45370	7922300	CHD	Allograft replacement of four or more coronary arteries
59423	7922y00	CHD	Other specified allograft replacement of coronary artery
48767	7922z00	CHD	Allograft replacement of coronary artery NOS
19402	7923.00	CHD	Prosthetic replacement of coronary artery
36011	7923.11	CHD	Prosthetic bypass of coronary artery
92419	7923000	CHD	Prosthetic replacement of one coronary artery
66664	7923100	CHD	Prosthetic replacement of two coronary arteries
66236	7923200	CHD	Prosthetic replacement of three coronary arteries
67761	7923300	CHD	Prosthetic replacement of four or more coronary arteries
19193	7923z00	CHD	Prosthetic replacement of coronary artery NOS
33461	7924.00	CHD	Revision of bypass for coronary artery
52938	7924000	CHD	Revision of bypass for one coronary artery
67554	7924100	CHD	Revision of bypass for two coronary arteries
31540	7924200	CHD	Revision of bypass for three coronary arteries
101569	7924300	CHD	Revision of bypass for four or more coronary arteries
63153	7924500	CHD	Revision of implantation of thoracic artery into heart
97953	7924y00	CHD	Other specified revision of bypass for coronary artery
57634	7924z00	CHD	Revision of bypass for coronary artery NOS
37682	7925.00	CHD	Connection of mammary artery to coronary artery
28837	7925.11	CHD	Creation of bypass from mammary artery to coronary artery
33718	7925000	CHD	Double anastomosis of mammary arteries to coronary arteries
48822	7925011	CHD	LIMA sequential anastomosis
92233	7925012	CHD	RIMA sequential anastomosis
31519	7925100	CHD	Double implant of mammary arteries into coronary arteries
44723	7925200	CHD	Single anast mammary art to left ant descend coronary art
51507	7925300	CHD	Single anastomosis of mammary artery to coronary artery NEC
22647	7925311	CHD	LIMA single anastomosis
68123	7925312	CHD	RIMA single anastomosis
68139	7925400	CHD	Single implantation of mammary artery into coronary artery
37719	7925y00	CHD	Connection of mammary artery to coronary artery OS
56990	7925z00	CHD	Connection of mammary artery to coronary artery NOS
96804	7926.00	CHD	Connection of other thoracic artery to coronary artery
62608	7926000	CHD	Double anastom thoracic arteries to coronary arteries NEC

67591	7926200	CHD	Single anastomosis of thoracic artery to coronary artery NEC
60753	7926300	CHD	Single implantation thoracic artery into coronary artery NEC
72780	7926z00	CHD	Connection of other thoracic artery to coronary artery NOS
47788	7927.00	CHD	Other open operations on coronary artery
18903	7927000	CHD	Repair of arteriovenous fistula of coronary artery
19164	7927100	CHD	Repair of aneurysm of coronary artery
61592	7927200	CHD	Transection of muscle bridge of coronary artery
48206	7927300	CHD	Transposition of coronary artery NEC
51702	7927400	CHD	Exploration of coronary artery
5744	7927500	CHD	Open angioplasty of coronary artery
95382	7927y00	CHD	Other specified other open operation on coronary artery
41757	7927z00	CHD	Other open operation on coronary artery NOS
2901	7928.00	CHD	Transluminal balloon angioplasty of coronary artery
5703	7928.11	CHD	Percutaneous balloon coronary angioplasty
18670	7928000	CHD	Percut transluminal balloon angioplasty one coronary artery
33735	7928100	CHD	Percut translum balloon angioplasty mult coronary arteries
42462	7928200	CHD	Percut translum balloon angioplasty bypass graft coronary a
86071	7928300	CHD	Percut translum cutting balloon angioplasty coronary artery
41547	7928y00	CHD	Transluminal balloon angioplasty of coronary artery OS
732	7928z00	CHD	Transluminal balloon angioplasty of coronary artery NOS
24888	7929.00	CHD	Other therapeutic transluminal operations on coronary artery
22828	7929000	CHD	Percutaneous transluminal laser coronary angioplasty
33650	7929100	CHD	Percut transluminal coronary thrombolysis with streptokinase
40996	7929111	CHD	Percut translum coronary thrombolytic therapy- streptokinase
66583	7929200	CHD	Percut translum inject therap subst to coronary artery NEC
19046	7929300	CHD	Rotary blade coronary angioplasty
8942	7929400	CHD	Insertion of coronary artery stent
42304	7929500	CHD	Insertion of drug-eluting coronary artery stent
93618	7929600	CHD	Percutaneous transluminal atherectomy of coronary artery
6182	7929y00	CHD	Other therapeutic transluminal op on coronary artery OS
31679	7929z00	CHD	Other therapeutic transluminal op on coronary artery NOS
34965	792A.00	CHD	Diagnostic transluminal operations on coronary artery
19681	792A000	CHD	Percutaneous transluminal angioscopy
43446	792A100	CHD	Intravascular ultrasound of coronary artery
56905	792Ay00	CHD	Diagnostic transluminal operation on coronary artery OS
61248	792Az00	CHD	Diagnostic transluminal operation on coronary artery NOS
33620	792B.00	CHD	Repair of coronary artery NEC
22020	792B000	CHD	Endarterectomy of coronary artery NEC
94783	792B100	CHD	Repair of rupture of coronary artery
93432	792B200	CHD	Repair of arteriovenous malformation of coronary artery
69247	792By00	CHD	Other specified repair of coronary artery
44585	792Bz00	CHD	Repair of coronary artery NOS
55598	792C.00	CHD	Other replacement of coronary artery
55092	792C000	CHD	Replacement of coronary arteries using multiple methods
93828	792Cy00	CHD	Other specified replacement of coronary artery
70755	792Cz00	CHD	Replacement of coronary artery NOS
34963	792D.00	CHD	Other bypass of coronary artery
3159	792Dy00	CHD	Other specified other bypass of coronary artery
33471	792Dz00	CHD	Other bypass of coronary artery NOS

31571	792y.00	CHD	Other specified operations on coronary artery
10603	792z.00	CHD	Coronary artery operations NOS
45960	8B27.00	CHD	Antianginal therapy
240	G3...00	CHD	Ischaemic heart disease
24783	G3...11	CHD	Arteriosclerotic heart disease
20416	G3...12	CHD	Atherosclerotic heart disease
1792	G3...13	CHD	IHD - Ischaemic heart disease
241	G30..00	CHD	Acute myocardial infarction
13566	G30..11	CHD	Attack - heart
2491	G30..12	CHD	Coronary thrombosis
30421	G30..13	CHD	Cardiac rupture following myocardial infarction (MI)
1204	G30..14	CHD	Heart attack
1677	G30..15	CHD	MI - acute myocardial infarction
13571	G30..16	CHD	Thrombosis - coronary
17689	G30..17	CHD	Silent myocardial infarction
12139	G300.00	CHD	Acute anterolateral infarction
5387	G301.00	CHD	Other specified anterior myocardial infarction
40429	G301000	CHD	Acute anteroapical infarction
17872	G301100	CHD	Acute anteroseptal infarction
14897	G301z00	CHD	Anterior myocardial infarction NOS
8935	G302.00	CHD	Acute inferolateral infarction
29643	G303.00	CHD	Acute inferoposterior infarction
23892	G304.00	CHD	Posterior myocardial infarction NOS
14898	G305.00	CHD	Lateral myocardial infarction NOS
63467	G306.00	CHD	True posterior myocardial infarction
3704	G307.00	CHD	Acute subendocardial infarction
9507	G307000	CHD	Acute non-Q wave infarction
10562	G307100	CHD	Acute non-ST segment elevation myocardial infarction
1678	G308.00	CHD	Inferior myocardial infarction NOS
30330	G309.00	CHD	Acute Q-wave infarct
17133	G30A.00	CHD	Mural thrombosis
32854	G30B.00	CHD	Acute posterolateral myocardial infarction
29758	G30X.00	CHD	Acute transmural myocardial infarction of unspecif site
12229	G30X000	CHD	Acute ST segment elevation myocardial infarction
34803	G30y.00	CHD	Other acute myocardial infarction
28736	G30y000	CHD	Acute atrial infarction
62626	G30y100	CHD	Acute papillary muscle infarction
41221	G30y200	CHD	Acute septal infarction
46017	G30yz00	CHD	Other acute myocardial infarction NOS
14658	G30z.00	CHD	Acute myocardial infarction NOS
27951	G31..00	CHD	Other acute and subacute ischaemic heart disease
23579	G310.00	CHD	Postmyocardial infarction syndrome
15661	G310.11	CHD	Dressler's syndrome
36523	G311.00	CHD	Preinfarction syndrome
4656	G311.11	CHD	Crescendo angina
39655	G311.12	CHD	Impending infarction
1431	G311.13	CHD	Unstable angina
19655	G311.14	CHD	Angina at rest
61072	G311000	CHD	Myocardial infarction aborted

55137	G311011	CHD	MI - myocardial infarction aborted
7347	G311100	CHD	Unstable angina
17307	G311200	CHD	Angina at rest
34328	G311300	CHD	Refractory angina
18118	G311400	CHD	Worsening angina
11983	G311500	CHD	Acute coronary syndrome
54251	G311z00	CHD	Preinfarction syndrome NOS
39449	G312.00	CHD	Coronary thrombosis not resulting in myocardial infarction
9413	G31y.00	CHD	Other acute and subacute ischaemic heart disease
9276	G31y000	CHD	Acute coronary insufficiency
68357	G31y100	CHD	Microinfarction of heart
39693	G31y200	CHD	Subendocardial ischaemia
21844	G31y300	CHD	Transient myocardial ischaemia
27977	G31yz00	CHD	Other acute and subacute ischaemic heart disease NOS
4017	G32..00	CHD	Old myocardial infarction
16408	G32..11	CHD	Healed myocardial infarction
17464	G32..12	CHD	Personal history of myocardial infarction
1430	G33..00	CHD	Angina pectoris
20095	G330.00	CHD	Angina decubitus
18125	G330000	CHD	Nocturnal angina
29902	G330z00	CHD	Angina decubitus NOS
12986	G331.00	CHD	Prinzmetal's angina
11048	G331.11	CHD	Variant angina pectoris
36854	G332.00	CHD	Coronary artery spasm
25842	G33z.00	CHD	Angina pectoris NOS
66388	G33z000	CHD	Status anginosus
54535	G33z100	CHD	Stenocardia
7696	G33z200	CHD	Syncope anginosa
1414	G33z300	CHD	Angina on effort
32450	G33z400	CHD	Ischaemic chest pain
9555	G33z500	CHD	Post infarct angina
26863	G33z600	CHD	New onset angina
12804	G33z700	CHD	Stable angina
28554	G33zz00	CHD	Angina pectoris NOS
28138	G34..00	CHD	Other chronic ischaemic heart disease
5413	G340.00	CHD	Coronary atherosclerosis
1655	G340.11	CHD	Triple vessel disease of the heart
1344	G340.12	CHD	Coronary artery disease
3999	G340000	CHD	Single coronary vessel disease
5254	G340100	CHD	Double coronary vessel disease
6331	G341.00	CHD	Aneurysm of heart
27484	G341.11	CHD	Cardiac aneurysm
2155	G341000	CHD	Ventricular cardiac aneurysm
67087	G341100	CHD	Other cardiac wall aneurysm
59193	G341200	CHD	Aneurysm of coronary vessels
91774	G341300	CHD	Acquired atrioventricular fistula of heart
41677	G341z00	CHD	Aneurysm of heart NOS
36609	G342.00	CHD	Atherosclerotic cardiovascular disease
7320	G343.00	CHD	Ischaemic cardiomyopathy

29421	G344.00	CHD	Silent myocardial ischaemia
34633	G34y.00	CHD	Other specified chronic ischaemic heart disease
24540	G34y000	CHD	Chronic coronary insufficiency
23078	G34y100	CHD	Chronic myocardial ischaemia
35713	G34yz00	CHD	Other specified chronic ischaemic heart disease NOS
15754	G34z.00	CHD	Other chronic ischaemic heart disease NOS
18889	G34z000	CHD	Asymptomatic coronary heart disease
18842	G35..00	CHD	Subsequent myocardial infarction
45809	G350.00	CHD	Subsequent myocardial infarction of anterior wall
38609	G351.00	CHD	Subsequent myocardial infarction of inferior wall
72562	G353.00	CHD	Subsequent myocardial infarction of other sites
46166	G35X.00	CHD	Subsequent myocardial infarction of unspecified site
36423	G36..00	CHD	Certain current complication follow acute myocardial infarct
24126	G360.00	CHD	Haemopericardium/current comp folow acut myocard infarct
23708	G361.00	CHD	Atrial septal defect/curr comp folow acut myocardal infarct
37657	G362.00	CHD	Ventric septal defect/curr comp fol acut myocardal infarctn
59189	G363.00	CHD	Ruptur cardiac wall w'out haemopericard/cur comp fol ac MI
59940	G364.00	CHD	Ruptur chordae tendinae/curr comp fol acute myocard infarct
69474	G365.00	CHD	Rupture papillary muscle/curr comp fol acute myocard infarct
29553	G366.00	CHD	Thrombosis atrium,auric append&vent/curr comp foll acute MI
8568	G37..00	CHD	Cardiac syndrome X
32272	G38..00	CHD	Postoperative myocardial infarction
46112	G380.00	CHD	Postoperative transmural myocardial infarction anterior wall
46276	G381.00	CHD	Postoperative transmural myocardial infarction inferior wall
41835	G384.00	CHD	Postoperative subendocardial myocardial infarction
68748	G38z.00	CHD	Postoperative myocardial infarction, unspecified
22383	G3y..00	CHD	Other specified ischaemic heart disease
1676	G3z..00	CHD	Ischaemic heart disease NOS
35119	G501.00	CHD	Post infarction pericarditis
39546	Gyu3000	CHD	[X]Other forms of angina pectoris
68401	Gyu3200	CHD	[X]Other forms of acute ischaemic heart disease
47637	Gyu3300	CHD	[X]Other forms of chronic ischaemic heart disease
96838	Gyu3400	CHD	[X]Acute transmural myocardial infarction of unspecif site
99991	Gyu3600	CHD	[X]Subsequent myocardial infarction of unspecified site
code	type	endpoint	description
430	ICD9	stroke	Subarachnoid hemorrhage
431	ICD9	stroke	Intracerebral hemorrhage
432	ICD9	stroke	Other and unspecified intracranial hemorrhage
432.9	ICD9	stroke	Hemorrhage, intracranial, NOS
433	ICD9	stroke	Occlusion and stenosis of precerebral arteries
433.0	ICD9	stroke	Occlusion and stenosis of basilar artery
433.1	ICD9	stroke	Occlusion and stenosis of carotid artery
433.2	ICD9	stroke	Occlusion and stenosis of vertebral artery
434	ICD9	stroke	Occlusion of cerebral arteries
434.0	ICD9	stroke	Cerebral thrombosis
434.00	ICD9	stroke	Cerebral thrombosis without cerebral infarction
434.01	ICD9	stroke	Cerebral thrombosis with cerebral infarction
434.1	ICD9	stroke	Cerebral embolism
434.10	ICD9	stroke	Cerebral embolism without cerebral infarction

434.11	ICD9	stroke	Cerebral embolism with cerebral infarction
434.9	ICD9	stroke	Cerebral artery occlusion, unspecified
435	ICD9	stroke	Transient cerebral ischemia
435.0	ICD9	stroke	Basilar artery syndrome
435.1	ICD9	stroke	Vertebral artery syndrome
435.2	ICD9	stroke	Subclavian steal syndrome
435.3	ICD9	stroke	Vertebrobasilar artery syndrome
435.9	ICD9	stroke	Transient ischemic attack, unspec.
436	ICD9	stroke	Acute but ill-defined cerebrovascular disease
437	ICD9	stroke	Other and ill-defined cerebrovascular disease
437.0	ICD9	stroke	Cerebral atherosclerosis
437.1	ICD9	stroke	Other generalized ischemic cerebrovascular disease
437.2	ICD9	stroke	Hypertensive encephalopathy
437.3	ICD9	stroke	Cerebral aneurysm nonruptured
437.4	ICD9	stroke	Cerebral arteritis
437.5	ICD9	stroke	Moyamoya disease
437.6	ICD9	stroke	Nonpyogenic thrombosis of intracranial venous sinus
437.7	ICD9	stroke	Transient global amnesia
438	ICD9	stroke	Late effects of cerebrovascular disease
438.0	ICD9	stroke	Cognitive deficits
438.1	ICD9	stroke	Speech and language deficits
438.10	ICD9	stroke	Speech and language deficits, unspecified
438.11	ICD9	stroke	Aphasia
438.12	ICD9	stroke	Dysphasia
438.19	ICD9	stroke	Other speech and language deficits
438.2	ICD9	stroke	Hemiplegia/hemiparesis
438.20	ICD9	stroke	Hemiplegia affecting unspecified side
438.21	ICD9	stroke	Hemiplegia affecting dominant side
438.22	ICD9	stroke	Hemiplegia affecting nondominant side
438.3	ICD9	stroke	Monoplegia of upper limb
438.4	ICD9	stroke	Monoplegia of lower limb
438.5	ICD9	stroke	Other paralytic syndrome
438.8	ICD9	stroke	Other late effects of cerebrovascular disease
438.81	ICD9	stroke	Apraxia cerebrovascular disease
438.82	ICD9	stroke	Dysphagia cerebrovascular disease
438.83	ICD9	stroke	Facial weakness
438.84	ICD9	stroke	Ataxia
438.85	ICD9	stroke	Vertigo
438.9	ICD9	stroke	CVA, late effect, unspec.
G45	ICD10	stroke	Transient cerebral ischaemic attacks and related syndromes
G45.0	ICD10	stroke	Vertebro-basilar artery syndrome
G45.1	ICD10	stroke	Carotid artery syndrome (hemispheric)
G45.2	ICD10	stroke	Multiple and bilateral precerebral artery syndromes
G45.4	ICD10	stroke	Transient global amnesia
G45.8	ICD10	stroke	Other transient cerebral ischaemic attacks and related syndromes
G45.9	ICD10	stroke	Transient cerebral ischaemic attack, unspecified
I60	ICD10	stroke	Subarachnoid haemorrhage
I60.0	ICD10	stroke	Subarachnoid haemorrhage from carotid siphon and bifurcation
I60.1	ICD10	stroke	Subarachnoid haemorrhage from middle cerebral artery

I60.2	ICD10	stroke	Subarachnoid haemorrhage from anterior communicating artery
I60.3	ICD10	stroke	Subarachnoid haemorrhage from posterior communicating artery
I60.4	ICD10	stroke	Subarachnoid haemorrhage from basilar artery
I60.5	ICD10	stroke	Subarachnoid haemorrhage from vertebral artery
I60.6	ICD10	stroke	Subarachnoid haemorrhage from other intracranial arteries
I60.7	ICD10	stroke	Subarachnoid haemorrhage from intracranial artery, unspecified
I60.8	ICD10	stroke	Other subarachnoid haemorrhage
I60.9	ICD10	stroke	Subarachnoid haemorrhage, unspecified
I61	ICD10	stroke	Intracerebral haemorrhage
I61.0	ICD10	stroke	Intracerebral haemorrhage in hemisphere, subcortical
I61.1	ICD10	stroke	Intracerebral haemorrhage in hemisphere, cortical
I61.2	ICD10	stroke	Intracerebral haemorrhage in hemisphere, unspecified
I61.3	ICD10	stroke	Intracerebral haemorrhage in brain stem
I61.4	ICD10	stroke	Intracerebral haemorrhage in cerebellum
I61.5	ICD10	stroke	Intracerebral haemorrhage, intraventricular
I61.6	ICD10	stroke	Intracerebral haemorrhage, multiple localized
I61.8	ICD10	stroke	Other intracerebral haemorrhage
I61.9	ICD10	stroke	Intracerebral haemorrhage, unspecified
I62	ICD10	stroke	Other nontraumatic intracranial haemorrhage
I62.0	ICD10	stroke	Subdural haemorrhage (acute)(nontraumatic)
I62.1	ICD10	stroke	Nontraumatic extradural haemorrhage
I62.9	ICD10	stroke	Intracranial haemorrhage (nontraumatic), unspecified
I63	ICD10	stroke	Cerebral infarction
I63.0	ICD10	stroke	Cerebral infarction due to thrombosis of precerebral arteries
I63.1	ICD10	stroke	Cerebral infarction due to embolism of precerebral arteries
I63.2	ICD10	stroke	Cerebral infarction due to unspecified occlusion or stenosis of precerebra arteries
I63.3	ICD10	stroke	Cerebral infarction due to thrombosis of cerebral arteries
I63.4	ICD10	stroke	Cerebral infarction due to embolism of cerebral arteries
I63.5	ICD10	stroke	Cerebral infarction due to unspecified occlusion or stenosis of cerebral arteries
I63.6	ICD10	stroke	Cerebral infarction due to cerebral venous thrombosis, nonpyogenic
I63.8	ICD10	stroke	Other cerebral infarction
I63.9	ICD10	stroke	Cerebral infarction, unspecified
I64	ICD10	stroke	Stroke, not specified as haemorrhage or infarction
I65	ICD10	stroke	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
I65.0	ICD10	stroke	Occlusion and stenosis of vertebral artery
I65.1	ICD10	stroke	Occlusion and stenosis of basilar artery
I65.2	ICD10	stroke	Occlusion and stenosis of carotid artery
I65.3	ICD10	stroke	Occlusion and stenosis of multiple and bilateral precerebral arteries
I65.8	ICD10	stroke	Occlusion and stenosis of other precerebral artery
I65.9	ICD10	stroke	Occlusion and stenosis of unspecified precerebral artery
I66	ICD10	stroke	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
I66.0	ICD10	stroke	Occlusion and stenosis of middle cerebral artery
I66.1	ICD10	stroke	Occlusion and stenosis of anterior cerebral artery
I66.2	ICD10	stroke	Occlusion and stenosis of posterior cerebral artery
I66.3	ICD10	stroke	Occlusion and stenosis of cerebellar arteries
I66.4	ICD10	stroke	Occlusion and stenosis of multiple and bilateral cerebral arteries

I66.8	ICD10	stroke	Occlusion and stenosis of other cerebral artery
I66.9	ICD10	stroke	Occlusion and stenosis of unspecified cerebral artery
I67.1	ICD10	stroke	Cerebral aneurysm, nonruptured
I67.2	ICD10	stroke	Cerebral atherosclerosis
I67.4	ICD10	stroke	Hypertensive encephalopathy
I67.5	ICD10	stroke	Moyamoya disease
I67.6	ICD10	stroke	Nonpyogenic thrombosis of intracranial venous system
I67.7	ICD10	stroke	Cerebral arteritis, not elsewhere classified
I67.8	ICD10	stroke	Other specified cerebrovascular diseases
I67.9	ICD10	stroke	Cerebrovascular disease, unspecified
I68.0	ICD10	stroke	Cerebral amyloid angiopathy
I68.2	ICD10	stroke	Cerebral arteritis in other diseases classified elsewhere
I68.8	ICD10	stroke	Other cerebrovascular disorders in diseases classified elsewhere
I69	ICD10	stroke	Sequelae of cerebrovascular disease
I69.0	ICD10	stroke	Sequelae of subarachnoid haemorrhage
I69.1	ICD10	stroke	Sequelae of intracerebral haemorrhage
I69.2	ICD10	stroke	Sequelae of other nontraumatic intracranial haemorrhage
I69.3	ICD10	stroke	Sequelae of cerebral infarction
I69.4	ICD10	stroke	Sequelae of stroke, not specified as haemorrhage or infarction
I69.8	ICD10	stroke	Sequelae of other and unspecified cerebrovascular diseases
medcode	readcode	endpoint	description
34135	14A7.00	stroke	H/O: CVA/stroke
6305	14A7.11	stroke	H/O: CVA
5871	14A7.12	stroke	H/O: stroke
13567	14AB.00	stroke	H/O: TIA
16554	14AF.00	stroke	H/O sub-arachnoid haemorrhage
66873	14AK.00	stroke	H/O: Stroke in last year
100639	1M4..00	stroke	Central post-stroke pain
18686	662e.00	stroke	Stroke/CVA annual review
10792	662M.00	stroke	Stroke monitoring
28914	662o.00	stroke	Haemorrhagic stroke monitoring
35916	7A20300	stroke	Endarterectomy and patch repair of carotid artery
12733	7A20311	stroke	Carotid endarterectomy and patch
2654	7A20400	stroke	Endarterectomy of carotid artery NEC
25910	7A22.00	stroke	Transluminal operations on carotid artery
29973	7A22000	stroke	Percutaneous transluminal angioplasty of carotid artery
2659	7A22100	stroke	Arteriography of carotid artery
68069	7A22200	stroke	Endovascular repair of carotid artery
47580	7A22300	stroke	Percutaneous transluminal insertion stent carotid artery
62661	7A22y00	stroke	Other specified transluminal operation on carotid artery
41703	7A22z00	stroke	Transluminal operation on carotid artery NOS
53999	7A23.00	stroke	Cerebral artery and circle of Willis aneurysm operations
5365	7A23.11	stroke	Cerebral artery aneurysm operations
50929	7A23.12	stroke	Circle of Willis aneurysm operations
45897	7A23000	stroke	Excision of aneurysm of cerebral artery
71034	7A23100	stroke	Excision of aneurysm of circle of Willis
10625	7A23200	stroke	Clipping of aneurysm of cerebral artery
45450	7A23300	stroke	Clipping of aneurysm of circle of Willis
37823	7A23400	stroke	Ligation of aneurysm of cerebral artery NEC

97937	7A23500	stroke	Ligation of aneurysm of circle of Willis NEC
58757	7A23600	stroke	Obliteration of aneurysm of cerebral artery NEC
94491	7A23700	stroke	Obliteration of aneurysm of circle of Willis NEC
30415	7A23800	stroke	Percutaneous coil embolisation of cerebral artery aneurysm
63903	7A23y00	stroke	Operation on cerebral artery/ circle of Willis aneurysm OS
50260	7A23z00	stroke	Operation on cerebral artery/ circle of Willis aneurysm NOS
55351	7P24200	stroke	Delivery of rehabilitation for stroke
13707	8HBJ.00	stroke	Stroke / transient ischaemic attack referral
56458	8HHM.00	stroke	Ref to multidisciplinary stroke function improvement service
18804	8HTQ.00	stroke	Referral to stroke clinic
32959	9N0p.00	stroke	Seen in stroke clinic
18687	9N4X.00	stroke	DNA - Did not attend stroke clinic
31218	9Om..00	stroke	Stroke/transient ischaemic attack monitoring administration
28753	9Om0.00	stroke	Stroke/transient ischaemic attack monitoring first letter
34245	9Om1.00	stroke	Stroke/transient ischaemic attack monitoring second letter
34375	9Om2.00	stroke	Stroke/transient ischaemic attack monitoring third letter
51465	9Om3.00	stroke	Stroke/transient ischaemic attack monitoring verbal invitati
89913	9Om4.00	stroke	Stroke/transient ischaemic attack monitoring telephone invte
54744	F11x200	stroke	Cerebral degeneration due to cerebrovascular disease
2418	G6...00	stroke	Cerebrovascular disease
1786	G60..00	stroke	Subarachnoid haemorrhage
29939	G600.00	stroke	Ruptured berry aneurysm
56007	G601.00	stroke	Subarachnoid haemorrhage from carotid siphon and bifurcation
19412	G602.00	stroke	Subarachnoid haemorrhage from middle cerebral artery
42331	G603.00	stroke	Subarachnoid haemorrhage from anterior communicating artery
9696	G604.00	stroke	Subarachnoid haemorrhage from posterior communicating artery
41910	G605.00	stroke	Subarachnoid haemorrhage from basilar artery
60692	G606.00	stroke	Subarachnoid haemorrhage from vertebral artery
17326	G60X.00	stroke	Subarachnoid haemorrh from intracranial artery, unspecif
23580	G60z.00	stroke	Subarachnoid haemorrhage NOS
5051	G61..00	stroke	Intracerebral haemorrhage
6960	G61..11	stroke	CVA - cerebrovascular accid due to intracerebral haemorrhage
18604	G61..12	stroke	Stroke due to intracerebral haemorrhage
31595	G610.00	stroke	Cortical haemorrhage
40338	G611.00	stroke	Internal capsule haemorrhage
46316	G612.00	stroke	Basal nucleus haemorrhage
13564	G613.00	stroke	Cerebellar haemorrhage
7912	G614.00	stroke	Pontine haemorrhage
62342	G615.00	stroke	Bulbar haemorrhage
30045	G616.00	stroke	External capsule haemorrhage
30202	G617.00	stroke	Intracerebral haemorrhage, intraventricular
57315	G618.00	stroke	Intracerebral haemorrhage, multiple localized
31060	G61X.00	stroke	Intracerebral haemorrhage in hemisphere, unspecified
28314	G61X000	stroke	Left sided intracerebral haemorrhage, unspecified
19201	G61X100	stroke	Right sided intracerebral haemorrhage, unspecified
3535	G61z.00	stroke	Intracerebral haemorrhage NOS
31805	G62..00	stroke	Other and unspecified intracranial haemorrhage
36178	G620.00	stroke	Extradural haemorrhage - nontraumatic
4273	G621.00	stroke	Subdural haemorrhage - nontraumatic

17734	G622.00	stroke	Subdural haematoma - nontraumatic
18912	G623.00	stroke	Subdural haemorrhage NOS
20284	G62z.00	stroke	Intracranial haemorrhage NOS
45781	G63..00	stroke	Precerebral arterial occlusion
57495	G63..11	stroke	Infarction - precerebral
63830	G63..12	stroke	Stenosis of precerebral arteries
32447	G630.00	stroke	Basilar artery occlusion
4240	G631.00	stroke	Carotid artery occlusion
2156	G631.11	stroke	Stenosis, carotid artery
4152	G631.12	stroke	Thrombosis, carotid artery
40847	G632.00	stroke	Vertebral artery occlusion
98642	G633.00	stroke	Multiple and bilateral precerebral arterial occlusion
2652	G634.00	stroke	Carotid artery stenosis
51326	G63y.00	stroke	Other precerebral artery occlusion
23671	G63y000	stroke	Cerebral infarct due to thrombosis of precerebral arteries
24446	G63y100	stroke	Cerebral infarction due to embolism of precerebral arteries
71585	G63z.00	stroke	Precerebral artery occlusion NOS
8837	G64..00	stroke	Cerebral arterial occlusion
5363	G64..11	stroke	CVA - cerebral artery occlusion
569	G64..12	stroke	Infarction - cerebral
6155	G64..13	stroke	Stroke due to cerebral arterial occlusion
16517	G640.00	stroke	Cerebral thrombosis
36717	G640000	stroke	Cerebral infarction due to thrombosis of cerebral arteries
15019	G641.00	stroke	Cerebral embolism
34758	G641.11	stroke	Cerebral embolus
27975	G641000	stroke	Cerebral infarction due to embolism of cerebral arteries
3149	G64z.00	stroke	Cerebral infarction NOS
15252	G64z.11	stroke	Brainstem infarction NOS
5602	G64z.12	stroke	Cerebellar infarction
25615	G64z000	stroke	Brainstem infarction
47642	G64z100	stroke	Wallenberg syndrome
5185	G64z111	stroke	Lateral medullary syndrome
9985	G64z200	stroke	Left sided cerebral infarction
10504	G64z300	stroke	Right sided cerebral infarction
26424	G64z400	stroke	Infarction of basal ganglia
504	G65..00	stroke	Transient cerebral ischaemia
3132	G65..11	stroke	Drop attack
1433	G65..12	stroke	Transient ischaemic attack
2417	G65..13	stroke	Vertebro-basilar insufficiency
23942	G650.00	stroke	Basilar artery syndrome
5268	G650.11	stroke	Insufficiency - basilar artery
33377	G651.00	stroke	Vertebral artery syndrome
21118	G651000	stroke	Vertebro-basilar artery syndrome
23465	G652.00	stroke	Subclavian steal syndrome
44765	G653.00	stroke	Carotid artery syndrome hemispheric
50594	G654.00	stroke	Multiple and bilateral precerebral artery syndromes
6489	G655.00	stroke	Transient global amnesia
10794	G656.00	stroke	Vertebrobasilar insufficiency
19354	G65y.00	stroke	Other transient cerebral ischaemia

1895	G65z.00	stroke	Transient cerebral ischaemia NOS
55247	G65z000	stroke	Impending cerebral ischaemia
16507	G65z100	stroke	Intermittent cerebral ischaemia
15788	G65zz00	stroke	Transient cerebral ischaemia NOS
1469	G66..00	stroke	Stroke and cerebrovascular accident unspecified
1298	G66..11	stroke	CVA unspecified
6253	G66..12	stroke	Stroke unspecified
6116	G66..13	stroke	CVA - Cerebrovascular accident unspecified
18689	G660.00	stroke	Middle cerebral artery syndrome
19280	G661.00	stroke	Anterior cerebral artery syndrome
19260	G662.00	stroke	Posterior cerebral artery syndrome
8443	G663.00	stroke	Brain stem stroke syndrome
17322	G664.00	stroke	Cerebellar stroke syndrome
33499	G665.00	stroke	Pure motor lacunar syndrome
51767	G666.00	stroke	Pure sensory lacunar syndrome
7780	G667.00	stroke	Left sided CVA
12833	G668.00	stroke	Right sided CVA
16956	G669.00	stroke	Cerebral palsy, not congenital or infantile, acute
13577	G67..00	stroke	Other cerebrovascular disease
11171	G670.00	stroke	Cerebral atherosclerosis
5184	G670.11	stroke	Precerebral atherosclerosis
40053	G671.00	stroke	Generalised ischaemic cerebrovascular disease NOS
70536	G671000	stroke	Acute cerebrovascular insufficiency NOS
24385	G671100	stroke	Chronic cerebral ischaemia
12555	G671z00	stroke	Generalised ischaemic cerebrovascular disease NOS
3979	G672.00	stroke	Hypertensive encephalopathy
31816	G672.11	stroke	Hypertensive crisis
4635	G673.00	stroke	Cerebral aneurysm, nonruptured
22018	G673000	stroke	Dissection of cerebral arteries, nonruptured
35059	G673100	stroke	Carotico-cavernous sinus fistula
12634	G673200	stroke	Carotid artery dissection
97122	G673300	stroke	Vertebral artery dissection
22400	G674.00	stroke	Cerebral arteritis
10189	G674000	stroke	Cerebral amyloid angiopathy
32310	G675.00	stroke	Moyamoya disease
37947	G676.00	stroke	Nonpyogenic venous sinus thrombosis
39344	G676000	stroke	Cereb infarct due cerebral venous thrombosis, nonpyogenic
31704	G677.00	stroke	Occlusion/stenosis cerebral arts not result cerebral infarct
51759	G677000	stroke	Occlusion and stenosis of middle cerebral artery
57527	G677100	stroke	Occlusion and stenosis of anterior cerebral artery
65770	G677200	stroke	Occlusion and stenosis of posterior cerebral artery
55602	G677300	stroke	Occlusion and stenosis of cerebellar arteries
71274	G677400	stroke	Occlusion+stenosis of multiple and bilat cerebral arteries
9943	G678.00	stroke	Cereb autosom dominant arteriop subcort infarcts leukoenceph
98188	G679.00	stroke	Small vessel cerebrovascular disease
101733	G67A.00	stroke	Cerebral vein thrombosis
34117	G67y.00	stroke	Other cerebrovascular disease OS
37493	G67z.00	stroke	Other cerebrovascular disease NOS
23361	G68..00	stroke	Late effects of cerebrovascular disease

44740	G680.00	stroke	Sequelae of subarachnoid haemorrhage
48149	G681.00	stroke	Sequelae of intracerebral haemorrhage
43451	G682.00	stroke	Sequelae of other nontraumatic intracranial haemorrhage
39403	G683.00	stroke	Sequelae of cerebral infarction
51138	G68W.00	stroke	Sequelae/other + unspecified cerebrovascular diseases
6228	G68X.00	stroke	Sequelae of stroke,not specfd as h'morrhage or infarction
40758	G6W..00	stroke	Cereb infarct due unsp occlus/stenos precerebr arteries
33543	G6X..00	stroke	Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
51311	G6y..00	stroke	Other specified cerebrovascular disease
10062	G6z..00	stroke	Cerebrovascular disease NOS
73901	Gyu6.00	stroke	[X]Cerebrovascular diseases
65745	Gyu6100	stroke	[X]Other subarachnoid haemorrhage
53810	Gyu6200	stroke	[X]Other intracerebral haemorrhage
91627	Gyu6300	stroke	[X]Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
53745	Gyu6400	stroke	[X]Other cerebral infarction
90572	Gyu6500	stroke	[X]Occlusion and stenosis of other precerebral arteries
92036	Gyu6600	stroke	[X]Occlusion and stenosis of other cerebral arteries
96630	Gyu6F00	stroke	[X]Intracerebral haemorrhage in hemisphere, unspecified
94482	Gyu6G00	stroke	[X]Cereb infarct due unsp occlus/stenos precerebr arteries
42248	ZLEP.00	stroke	Discharge from stroke serv
19348	ZV12511	stroke	[V]Personal history of stroke
7138	ZV12512	stroke	[V]Personal history of cerebrovascular accident (CVA)
code	type	endpoint	description
G45	ICD-10	TIA	Transient cerebral ischaemic attacks and related syndromes (excl: neonatal cerebral ischaemia (P91.0))
G45.0	ICD-10	TIA	Vertebro-basilar artery syndrome
G45.1	ICD-10	TIA	Carotid artery syndrome (hemispheric)
G45.2	ICD-10	TIA	Multiple and bilateral precerebral artery syndromes
G45.3	ICD-10	TIA	Amaurosis fugax
G45.4	ICD-10	TIA	Transient global amnesia excl. amnesia NOS (R41.3)
G45.8	ICD-10	TIA	Other transient cerebral ischaemic attacks and related syndromes
G45.9	ICD-10	TIA	Transient cerebral ischaemic attack, unspecified (Spasm of cerebral artery Transient cerebral ischaemia NOS)

medcode	**readcode**	**desc**
21844	G31y300	Transient myocardial ischaemia
55878	Q494.00	Transient myocardial ischaemia of newborn
102326	1JK..00	Suspected transient ischaemic attack
1433	G65..12	Transient ischaemic attack
504	G65..00	Transient cerebral ischaemia
89913	9Om4.00	Stroke/transient ischaemic attack monitoring telephone invte
28753	9Om0.00	Stroke/transient ischaemic attack monitoring first letter
34375	9Om2.00	Stroke/transient ischaemic attack monitoring third letter
15788	G65zz00	Transient cerebral ischaemia NOS
19354	G65y.00	Other transient cerebral ischaemia
13707	8HBJ.00	Stroke / transient ischaemic attack referral
101251	ZV12D00	[V]Personal history of transient ischaemic attack
100015	8CRB.00	Transient ischaemic attack clinical management plan
31218	9Om..00	Stroke/transient ischaemic attack monitoring administration
34245	9Om1.00	Stroke/transient ischaemic attack monitoring second letter
51465	9Om3.00	Stroke/transient ischaemic attack monitoring verbal invitati

1895	G65z.00	Transient cerebral ischaemia NOS
42720	F580200	Transient ischaemic deafness
2417	G65..13	Vertebro-basilar insufficiency
10794	G656.00	Vertebrobasilar insufficiency
21118	G651000	Vertebro-basilar artery syndrome
44765	G653.00	Carotid artery syndrome hemispheric
51326	G63y.00	Other precerebral artery occlusion
71585	G63z.00	Precerebral artery occlusion NOS
50594	G654.00	Multiple and bilateral precerebral artery syndromes
1195	F423600	Amaurosis fugax
28278	1B1S.00	Transient global amnesia
19004	Z7CE711	TGA - Transient global amnesia
18996	Z7CE700	Transient global amnesia
6489	G655.00	Transient global amnesia

References

4 WHO. (2011). WHO factsheet No.37: CVD. Retrieved from http://www.who.int/mediacentre/factsheets/fs317/en/index.html

5 Department of Health. Putting prevention first—vascular checks: risk assessment andmanagement.DoH,2008:15.

6 National Institute for Health and Clinical Excellence. (2010). Lipid Modification: Cardiovascular risk assessment and the modifications of blood lipids for the primary and secondary prevention of cardiovascular disease. Guideline 67. *London: National Institute for Health and Clinical Excellence*., *2008*(March). Retrieved from http://www.nice.org.uk/nicemedia/pdf/CG67NICEguideline.pdf

7 Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation 2002;106:3140-1.

8 Graham I, Atar D, Borch-Johnsen K, Boysen G, Burell G, Cifkova R, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary. EurHeart J 2007;28:2375-414.

9 Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles.Am Heart J 1991;121(1 pt 2):293-8.

10 Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds risk score. JAMA 2007;297:611-9.

11 Conroy RM, Pyörälä K, Fitzgerald AP,SansS,Menotti A, de Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003;24:987-1003.

12 WoodwardM,Brindle P, Tunstall-PedoeH. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score fromthe Scottish HeartHealthExtendedCohort (SHHEC). Heart 2007;93:172-6.

13 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J,Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England andWales: prospective derivation and validation of QRISK2. BMJ 2008;336:1475-82, doi:10.1136/bmj.39609.449676.25.

14 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart 2008;94:34-9.

15 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J,MayM, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007;335:136, doi:10.1136/bmj.39261.471806.55.

16 British Cardiac Society, British Hypertension Society, Diabetes UK, HEART UK, Primary Care Cardiovascular Society, T. S. A. (2005). JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart (British Cardiac Society)*, *91 Suppl 5*, v1-52. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1876394&tool=pmcentrez&rendertype=abstract

17 Lloyd-JonesDM, LarsonMG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. Lancet 1999;353:89-92.

18 Asia Pacific Cohort Studies Collaboration. Cardiovascular risk prediction tools for populations in Asia. J Epidemiol Community Health 2007;61:115-21.

19 Hippisley-Cox, J., Coupland, C., Robson, J., & Brindle, P. (2010). Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *British Medical Journal*, *341*(dec09 1), c6624. BMJ Publishing Group Ltd. Retrieved from http://www.bmj.com/cgi/doi/10.1136/bmj.c6624

20 Elward K, Simpson R,Mendy P. Improving cardiovascular risk reduction for primary prevention-utility of lifetime risk assessment. Postgrad Med J 2010;122:192-9.

21 PencinaMJ, D'Agostino RB, Sr, LarsonMG,Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. Circulation 2009;119:3078-84.

22 Lloyd-JonesDM, Leip EP, Larson MG,D'Agostino RB,Beiser A, Wilson PWF, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 2006;113:791-8.

23 Berger JS, Jordan CO, Lloyd-Jones D, Blumenthal RS. Screening for cardiovascular risk in asymptomatic patients. JAmColl Cardiol 2010;55:1169-77.

24 Brindle P, Emberson J, Lampe F,WalkerM,Whincup P, Fahey T, et al. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study.BMJ 2003;327:1-6

25 Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing withmissing predictor values when applying clinical prediction models. Clin Chem 2009;55:994-1001.

26 Schafer JL. Multiple imputation: a primer. StatMethodsMed Res 1999;8:3-15.

27 Gail M, Pfeiffer R.Onevaluatingmodels of absolute risk. Biostatistics 2005;6:227-39.

28 Erika Graf CSWSMS. Assessment and comparison of prognostic classification schemes for survival data. StatMed1999;18:2529-45.

29 Royston P, SauerbreiW. A newmeasure of prognostic separation in survival data. StatMed 2004;23:723-48.

30 Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. doi:10.1016/j.msea.2008.11.058

31 Poli, R., & Langdon, W. (2008). A field guide to genetic programming, (March). Retrieved from http://books.google.com/books?hl=en&lr=&id=3PBrqNK5fFQC&oi=fnd&pg=PA1&dq=A+Field+Guide+to+Genetic+Programming&ots=fsvws7lZtN&sig=b4K3UFs0NmDz6bOPMWorsqKU8DA

32 Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, ISBN 0-262-11170-5.

33 W. Cai, A. Pacheco-Vega, M. Sen, and K. T. Yang. Heat transfer correlations by symbolic regression. International Journal of Heat and Mass Transfer, 49(23-24):4352–4359, November 2006.

34 S. M. Gustafson, E. K. Burke, and N. Krasnogor. On improving genetic programming for symbolic regression. In D. Corne, et al., editors, Proceedings of the 2005 IEEE Congress on Evolutionary Computation, volume 1, pages 912–919, Edinburgh, UK, 2-5 September 2005. IEEE Press. ISBN 0-7803-9363-5.

35 M. Keijzer. Scaled symbolic regression. Genetic Programming and Evolvable Machines, 5(3):259–269, September 2004. ISSN 1389-2576.

36 T. L. Lew, A. B. Spencer, F. Scarpa, K. Worden, A. Rutherford, and F. Hemez. Iden- tification of response surface models using genetic programming. Mechanical Systems and Signal Processing, 20(8):1819–1831, November 2006.

37 Liew, S. M., Doust, J., & Glasziou, P. (2011). Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart (British Cardiac Society)*, *97*(9), 689–97. doi:10.1136/hrt.2010.220442

Appendix G

Run statistics: CPRD experiments

The full range of run statistics for the 25 GP runs in the CPRD experiments in chapter 7

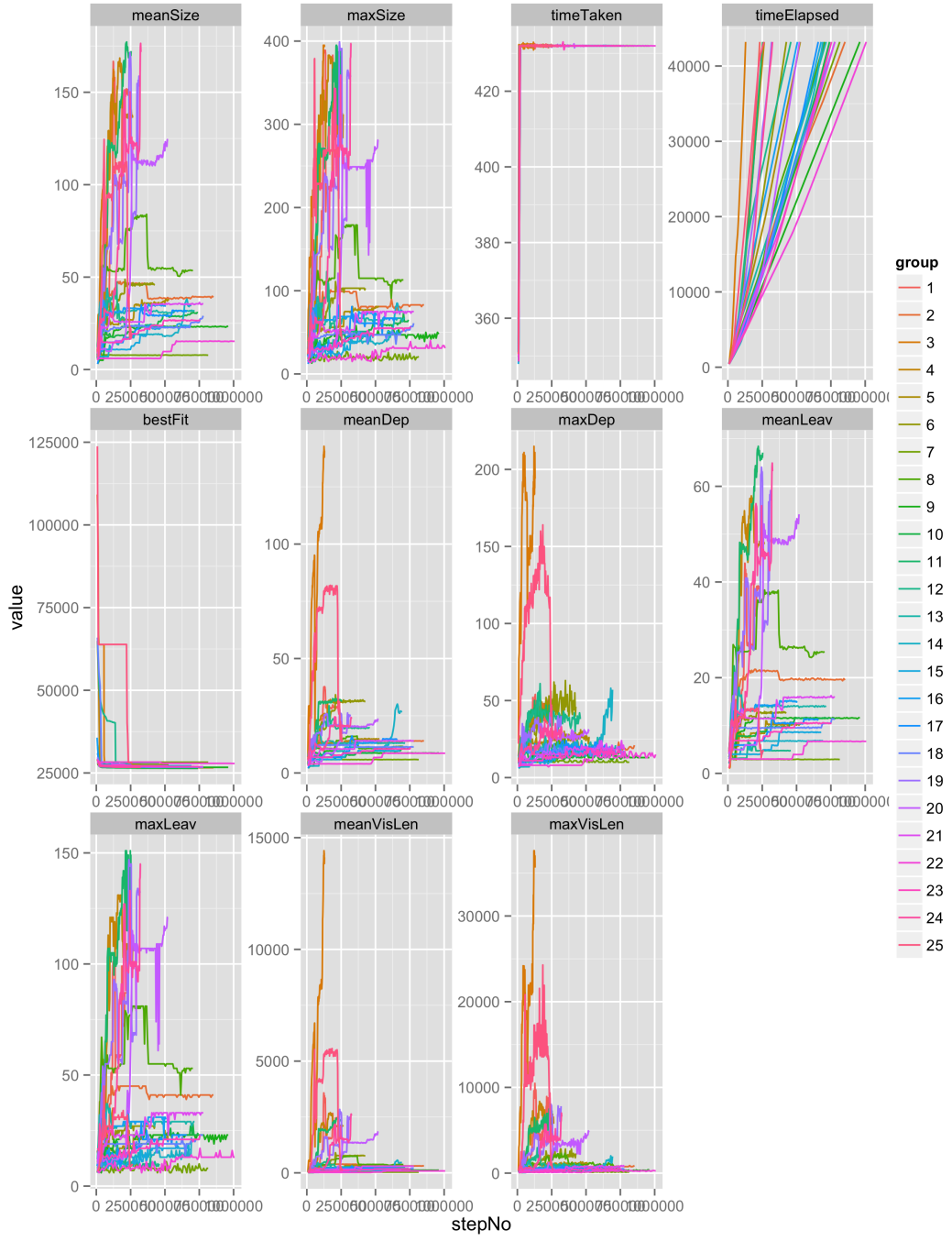
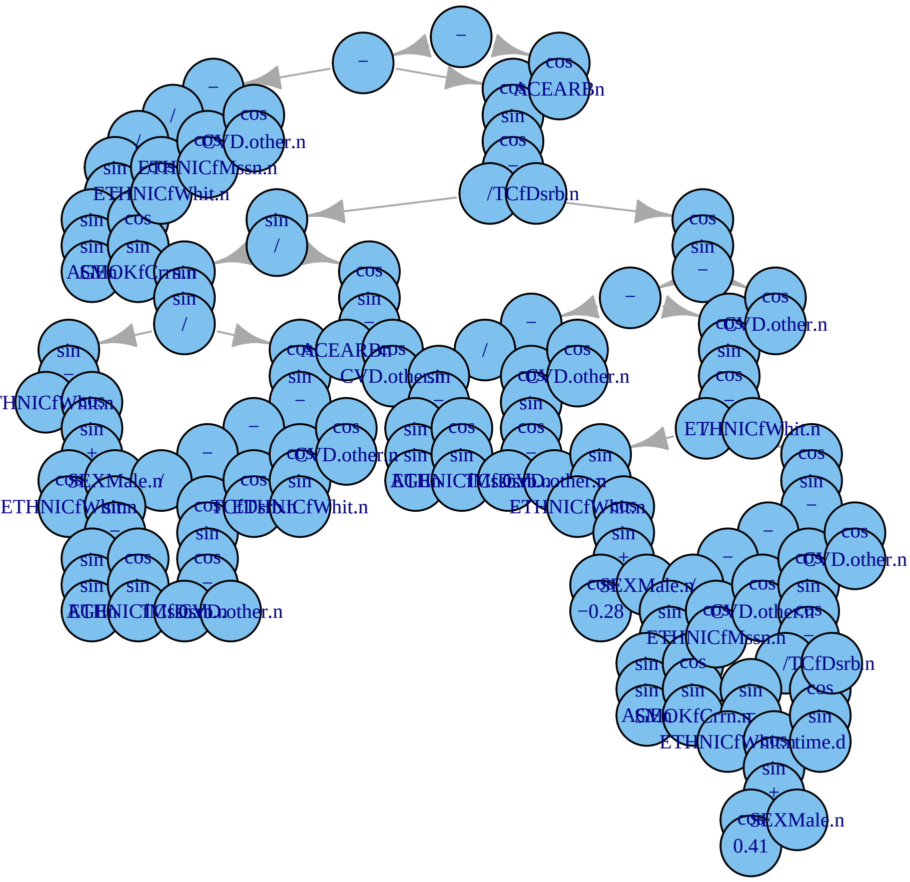


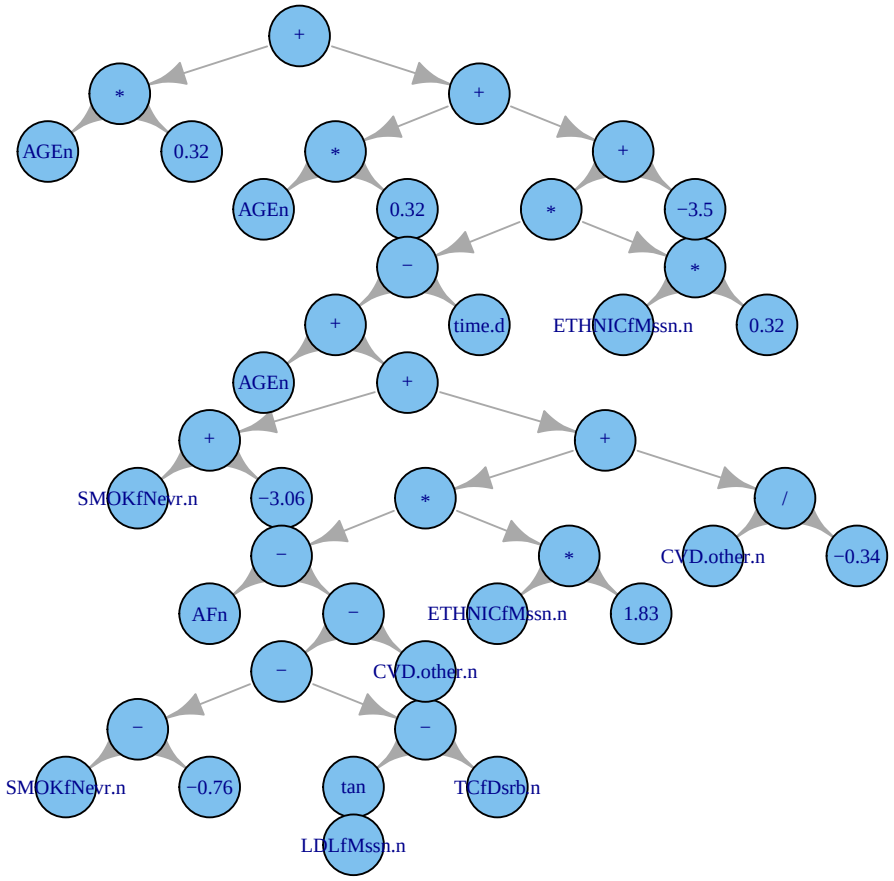
Figure G.1: The full range of runs statisitcs for the 25 SSOGP runs in the CPRD experiments in chapter 7.

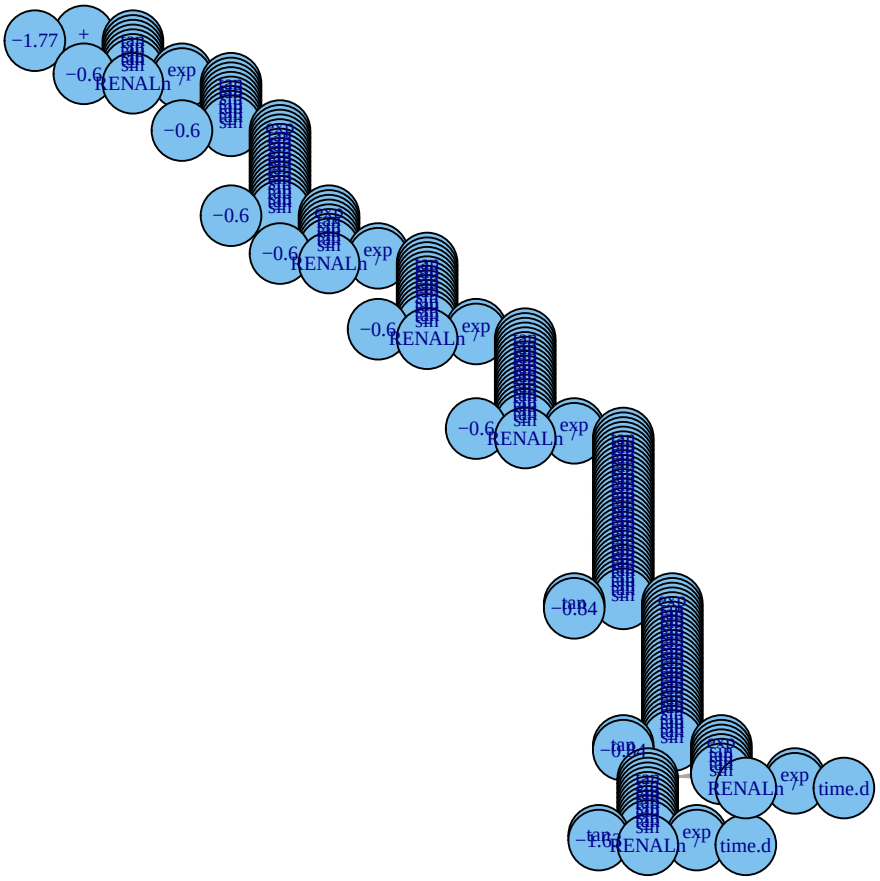
Appendix H

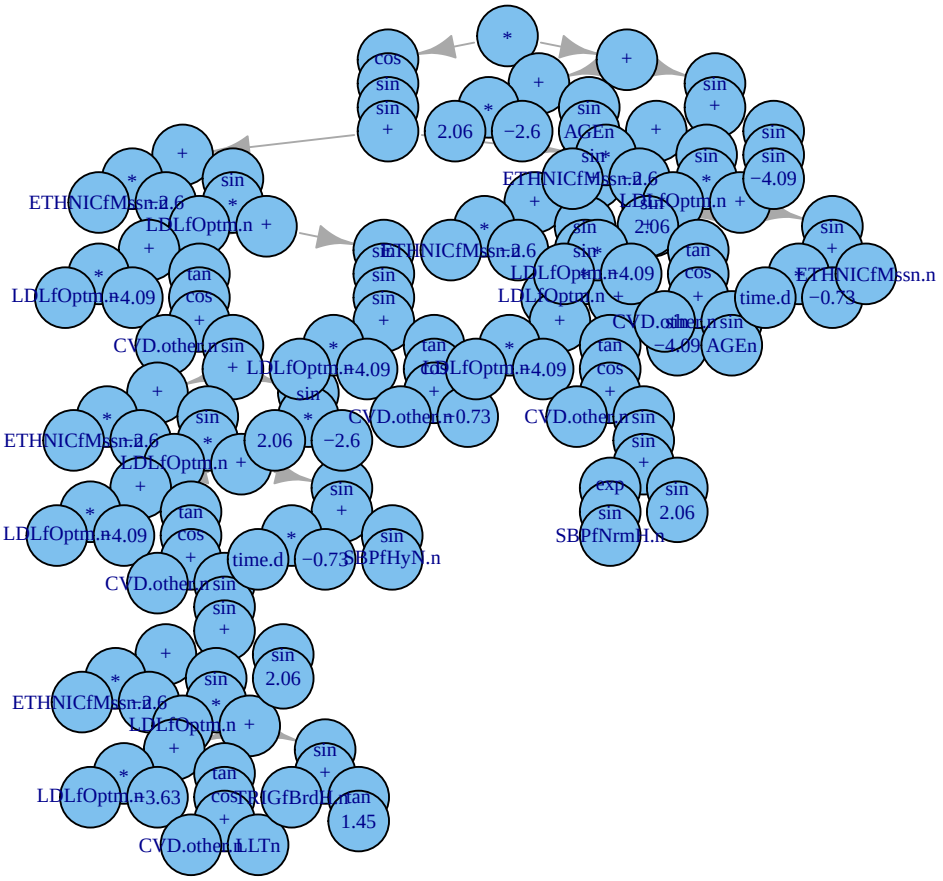
Final Models: CPRD experiments

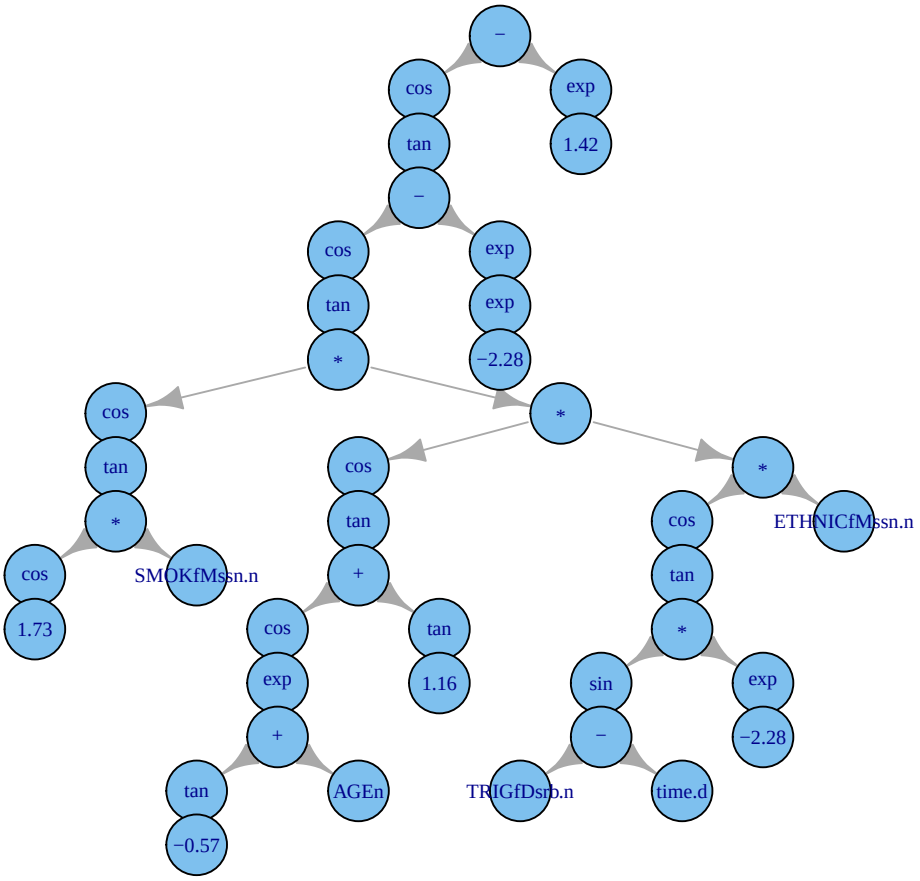
The final 25 models developed by SSOGP the CPRD experiments in chapter 7, presented as a binary trees

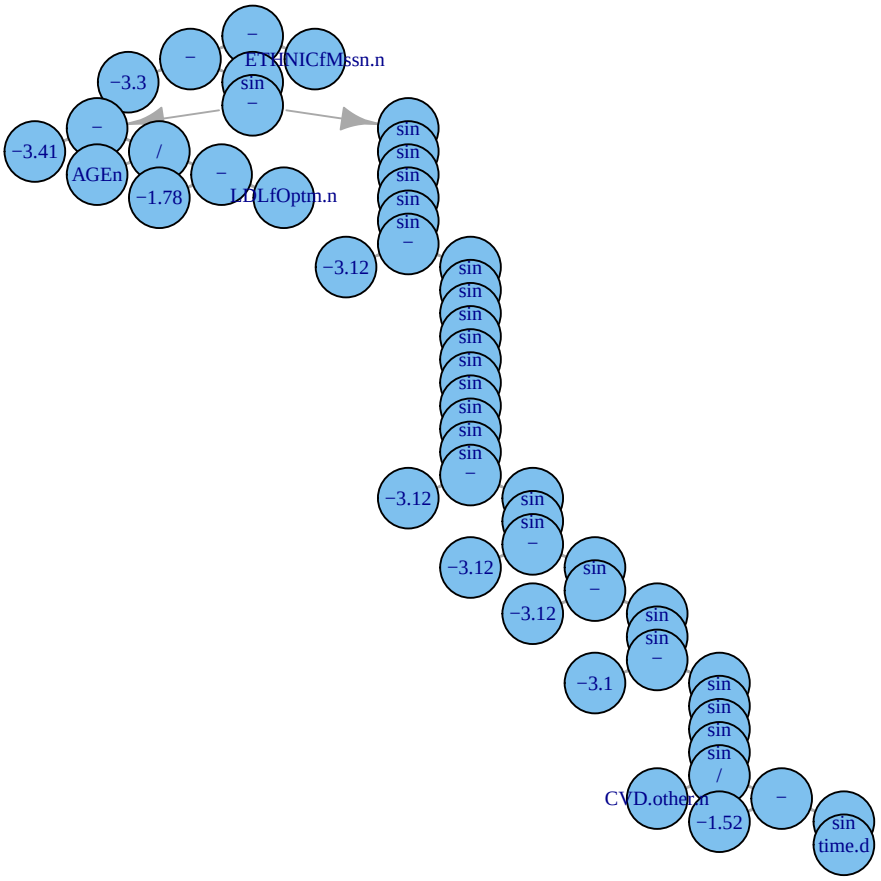

/TCfDsrb.n
ETHNICfWhit.n
CVD.other.n
SEXMale.n
−0.28
0.41

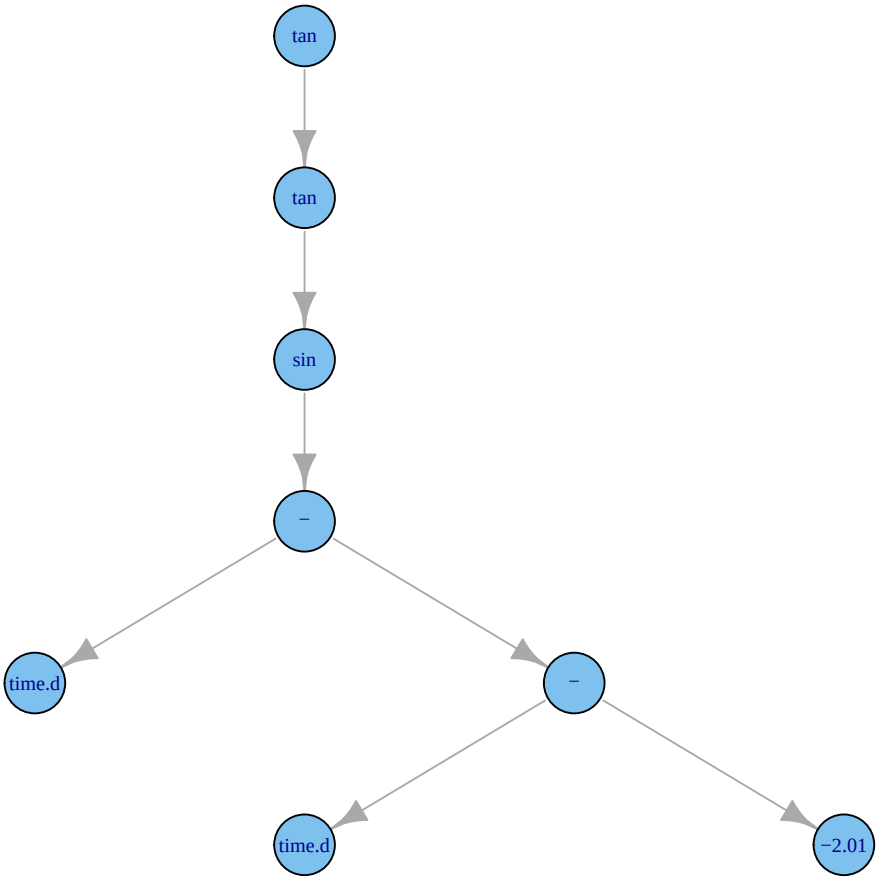

+
*
+
AGEn
0.32
*
+
AGEn
0.32
*
−3.5
−
*
+
time.d
ETHNICfMssn.n
0.32
AGEn
+
+
+
+
SMOKfNevr.n
−3.06
*
/
−
*
CVD.other.n
−0.34
AFn
−
ETHNICfMssn.n
1.83
−
CVD.other.n
−
−
SMOKfNevr.n
−0.76
tan
TCfDsrb.n
LDLfMssn.n

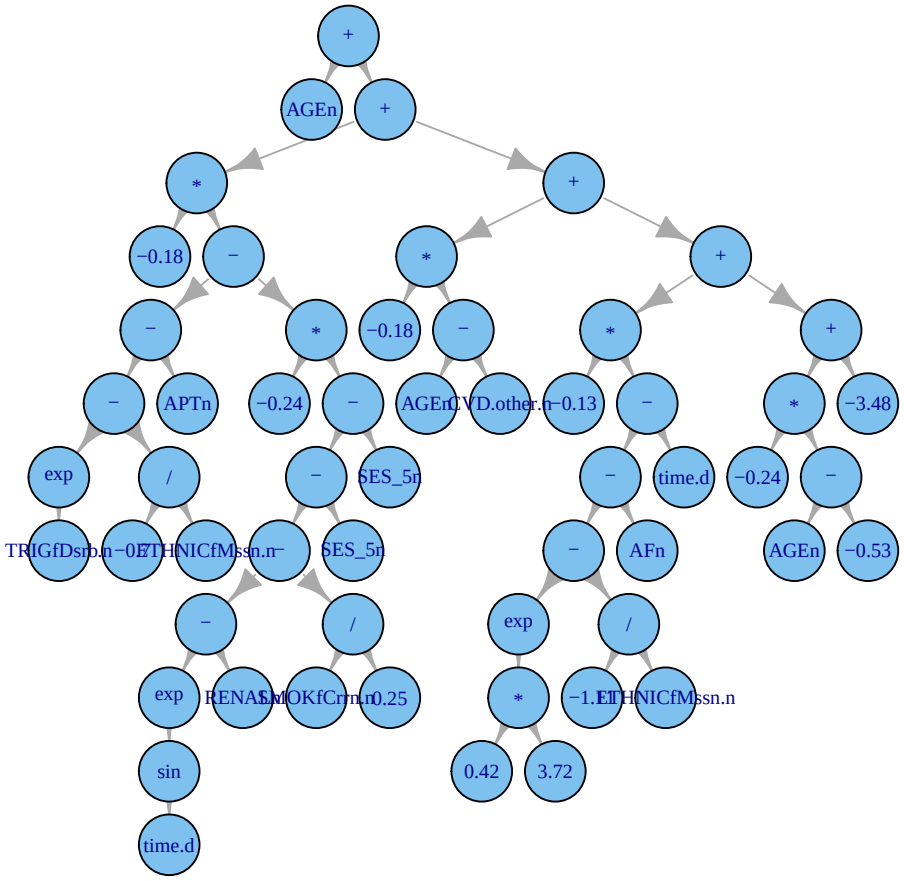

−1.77
+
−0.6
RENALn
exp
−0.6
−0.6
−0.6
RENALn
exp
−0.6
RENALn
exp
−0.6
RENALn
exp
−0.84
time.d
RENALn
exp
time.d
RENALn
exp

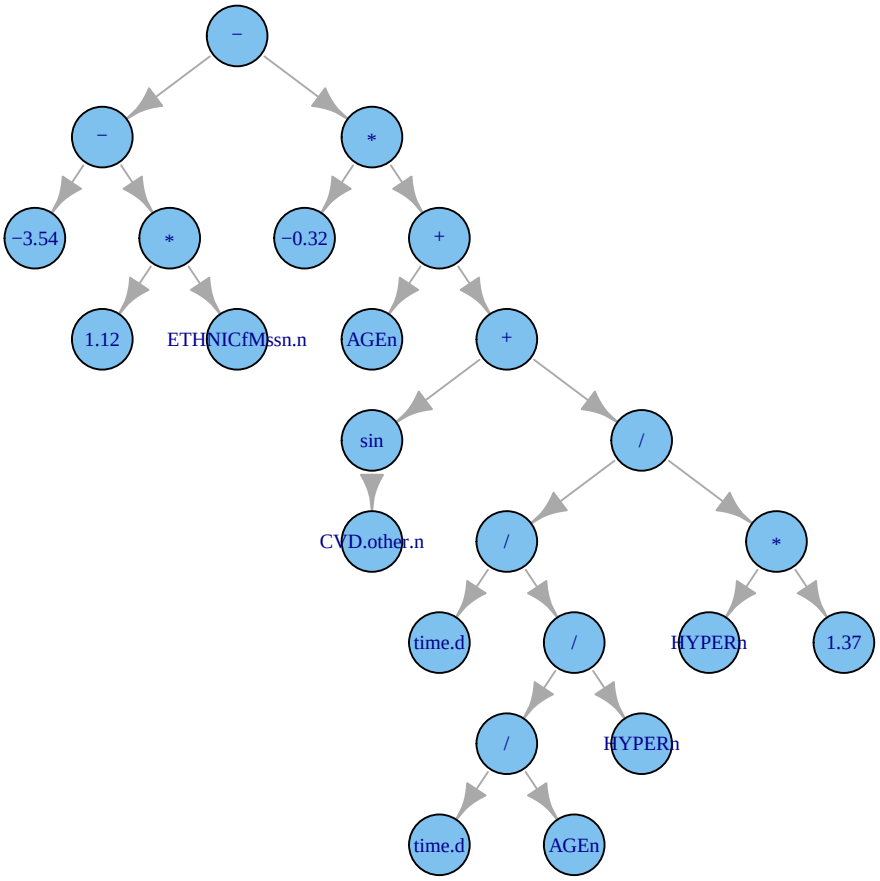

*
cos
sin
sin
+
+
+
2.06
*
−2.6
sin
AGEn
+
sin
+
sin
sin
−4.09
ETHNICfMssn.n
LDLfOptrm.n
2.06
sin
sin
tan
cos
+
time.d
−0.73
ETHNICfMssn.n
LDLfOptrm.n
4.09
CVD.otherLn
−4.09
AGEn
CVD.otherLn
0.73
SBPfNrmH.n
2.06
time.d
−0.73
SBPfHyN.n
LDLfOptrm.n
3.63
tan
1.45

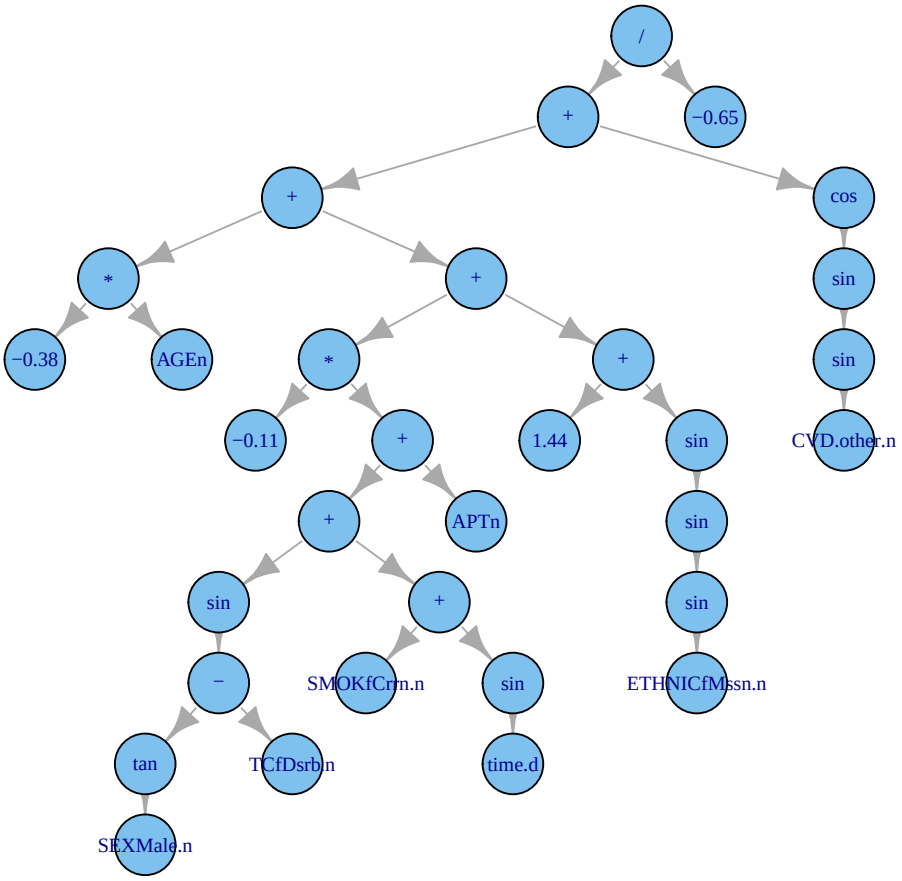

−
cos
exp
tan
1.42
−
cos
exp
tan
exp
*
−2.28
cos
*
tan
cos
*
*
cos
SMOKfMssn.n
tan
+
cos
tan
ETHNICfMssn.n
1.73
cos
tan
*
exp
1.16
sin
exp
+
−
−2.28
tan
AGEn
TRIGfDsrb.n
time.d
−0.57

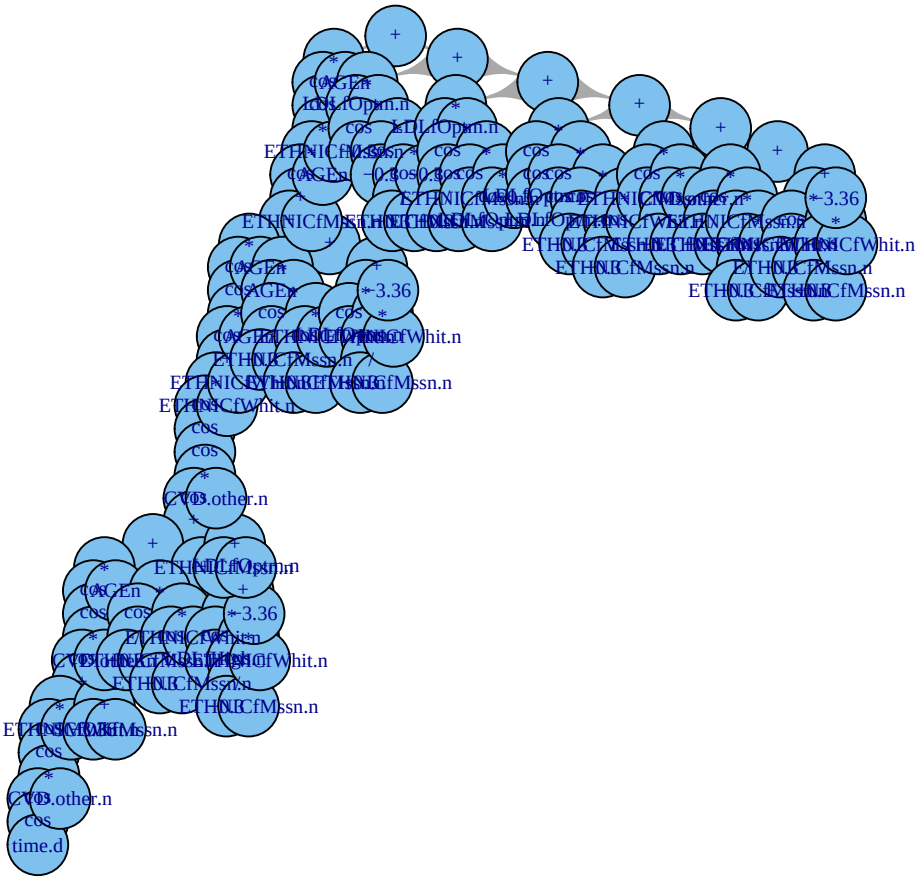

−3.3
−3.41
AGEn
−1.78
LDLfOptm.n
−3.12
−3.12
−3.12
−3.12
−3.1
CVD.other.n
−1.52
time.d

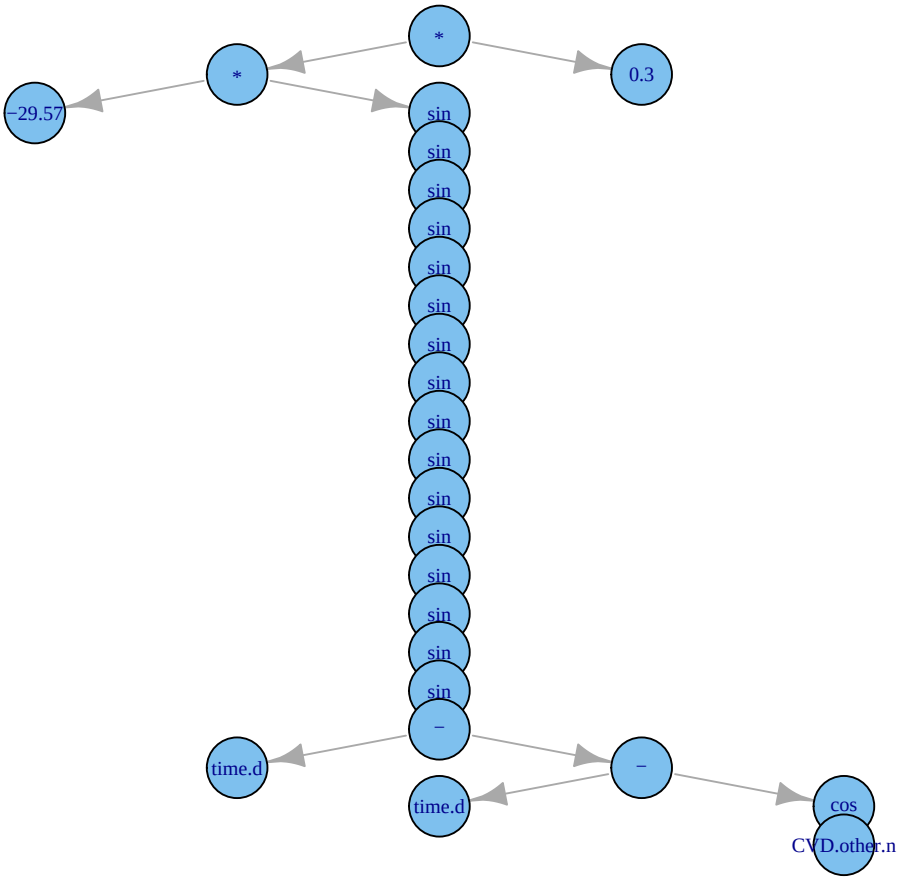

tan
tan
sin
−
time.d
−
time.d
−2.01

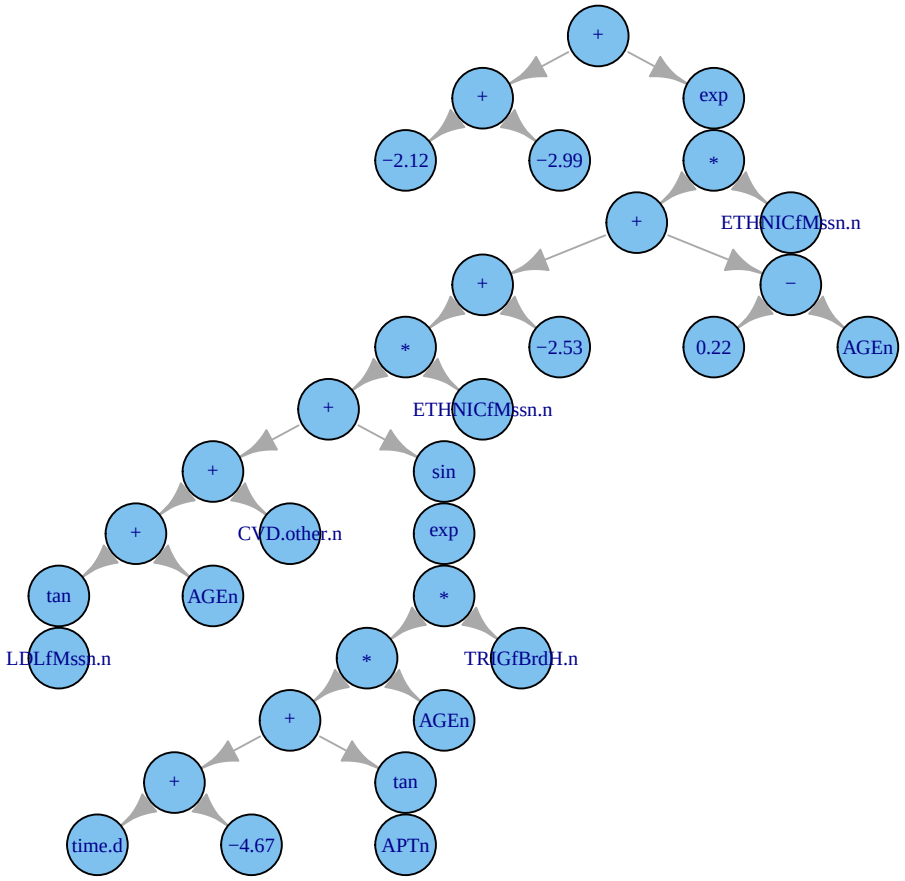

+
AGEn
+
*
+
−0.18
−
*
+
−
*
−0.18
−
*
+
−
APTn
−0.24
−
AGEn
CVD.other.n
−0.13
−
*
−3.48
exp
/
−
SES_5n
−
time.d
−0.24
−
TRIGfDsrb.n
−0.
ETHNICfMssn.n
−
SES_5n
−
AFn
AGEn
−0.53
−
/
exp
/
exp
*
−1.1
ETHNICfMssn.n
0.25
sin
0.42
3.72
time.d

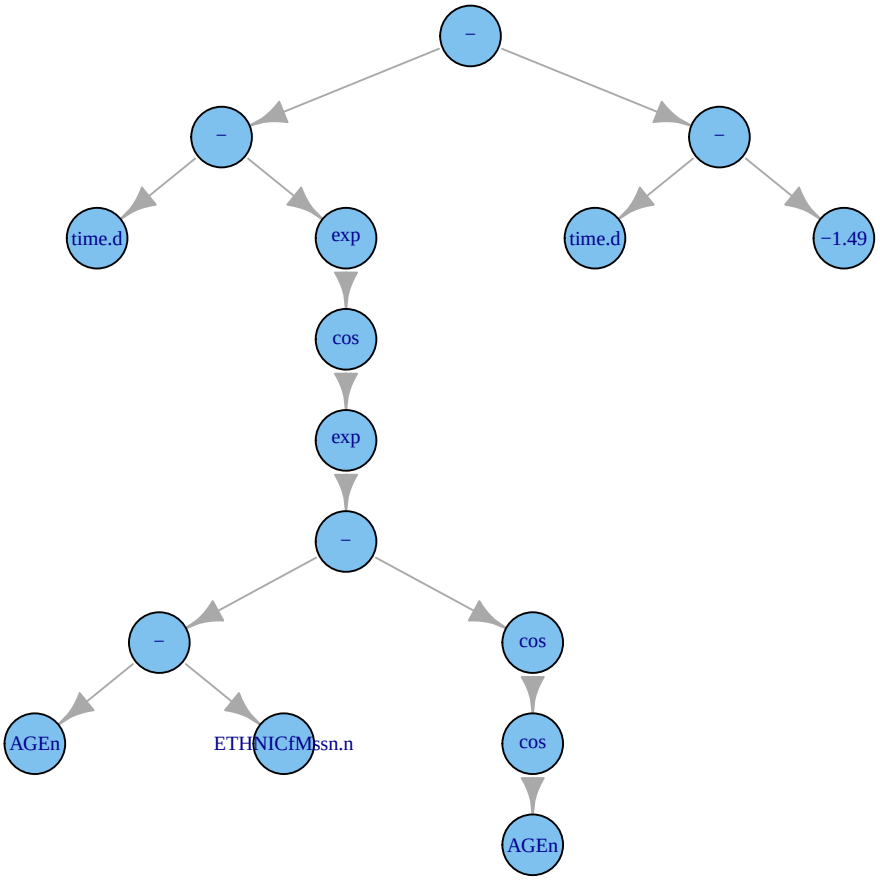

−
−
*
−3.54
*
−0.32
+
1.12
ETHNICfMssn.n
AGEn
+
sin
/
CVD.other.n
/
*
time.d
/
HYPERn
1.37
/
HYPERn
time.d
AGEn

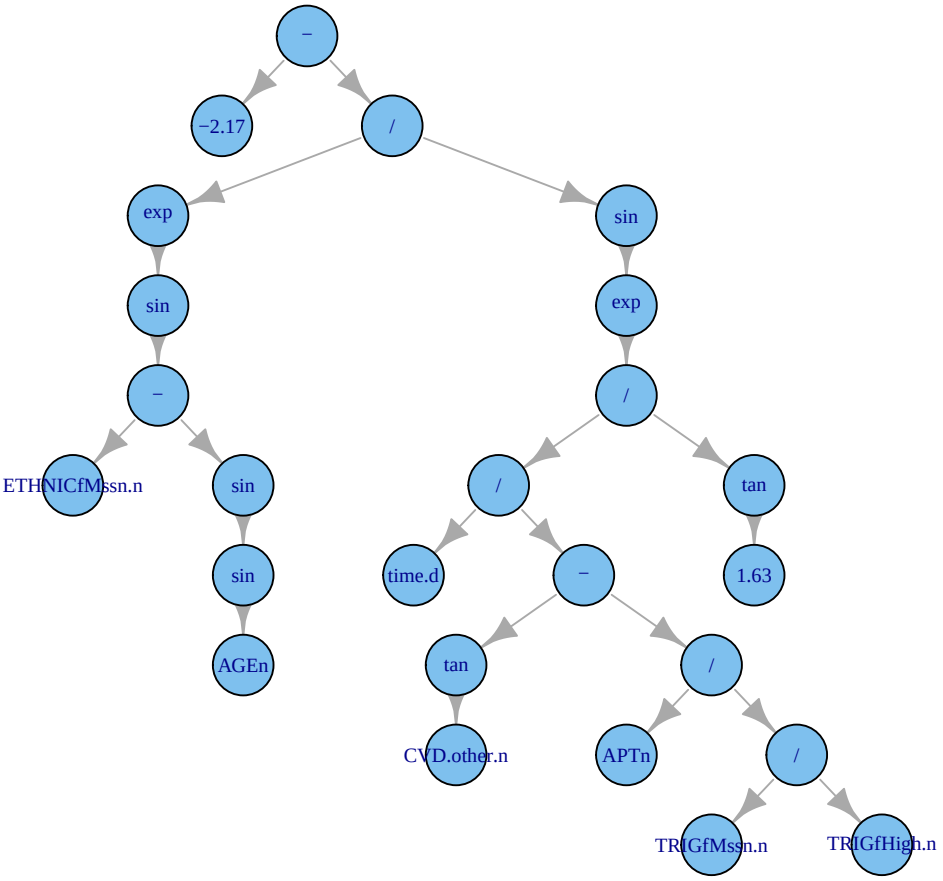

/
+
−0.65
+
cos
*
+
sin
−0.38
AGEn
*
+
sin
−0.11
+
1.44
sin
CVD.other.n
+
APTn
sin
sin
+
sin
−
SMOKfCrm.n
sin
ETHNICfMssn.n
tan
TCfDsrb.n
time.d
SEXMale.n

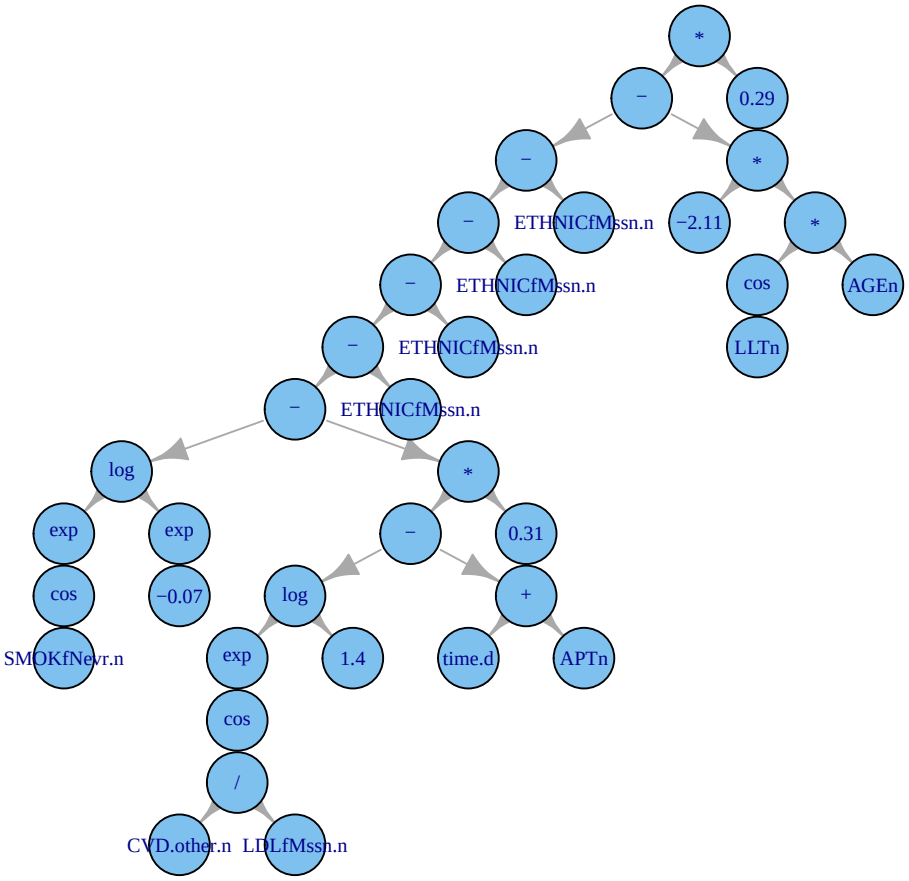

+
*
cos
-3.36
time.d

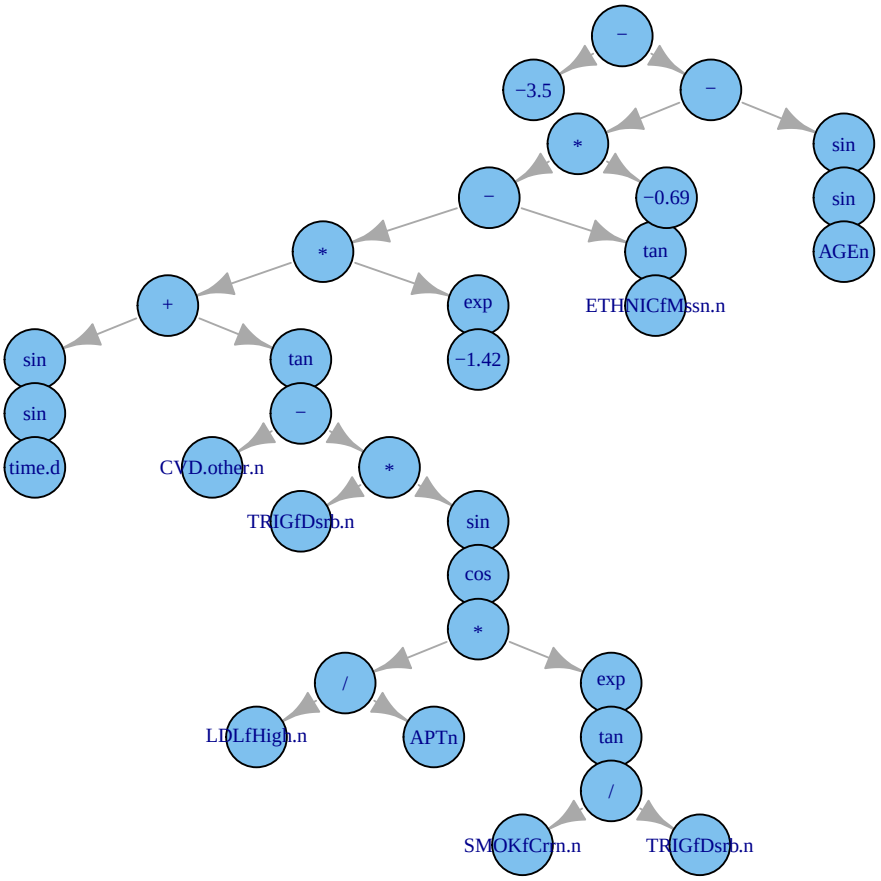

*
*
0.3
−29.57
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
sin
−
time.d
−
time.d
cos
CVD.other.n

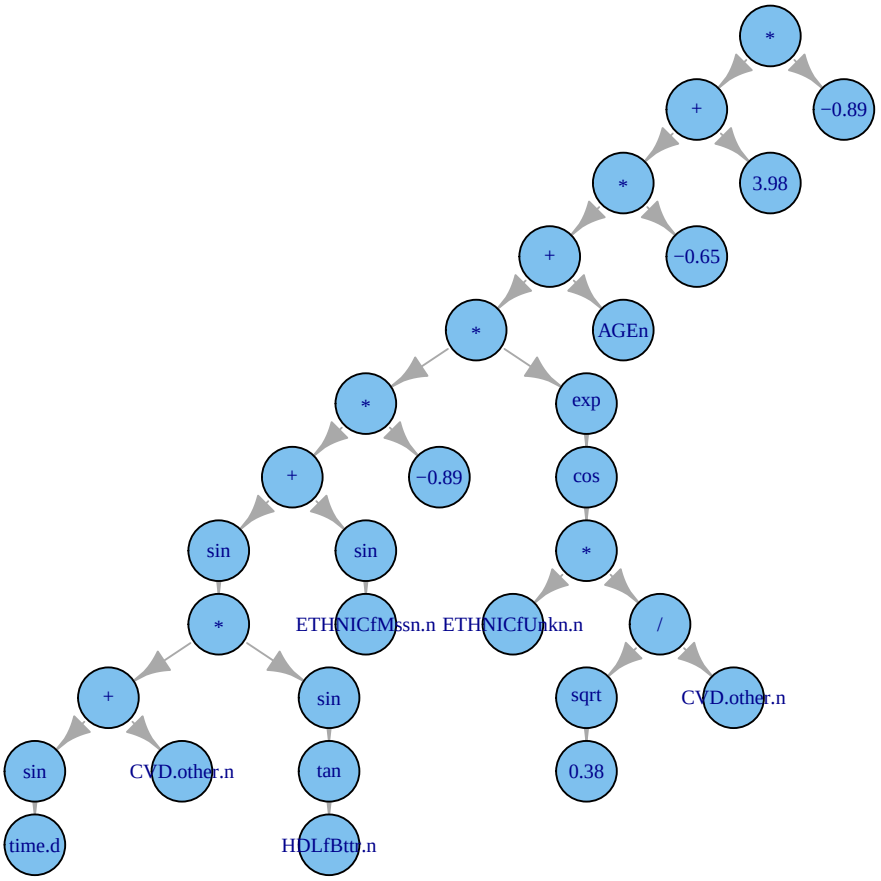

+
+
exp
−2.12
−2.99
*
+
ETHNICfMssn.n
+
−
*
−2.53
0.22
AGEn
+
ETHNICfMssn.n
+
sin
+
CVD.other.n
exp
tan
AGEn
*
LDLfMssn.n
*
TRIGfBrdH.n
+
AGEn
+
tan
time.d
−4.67
APTn

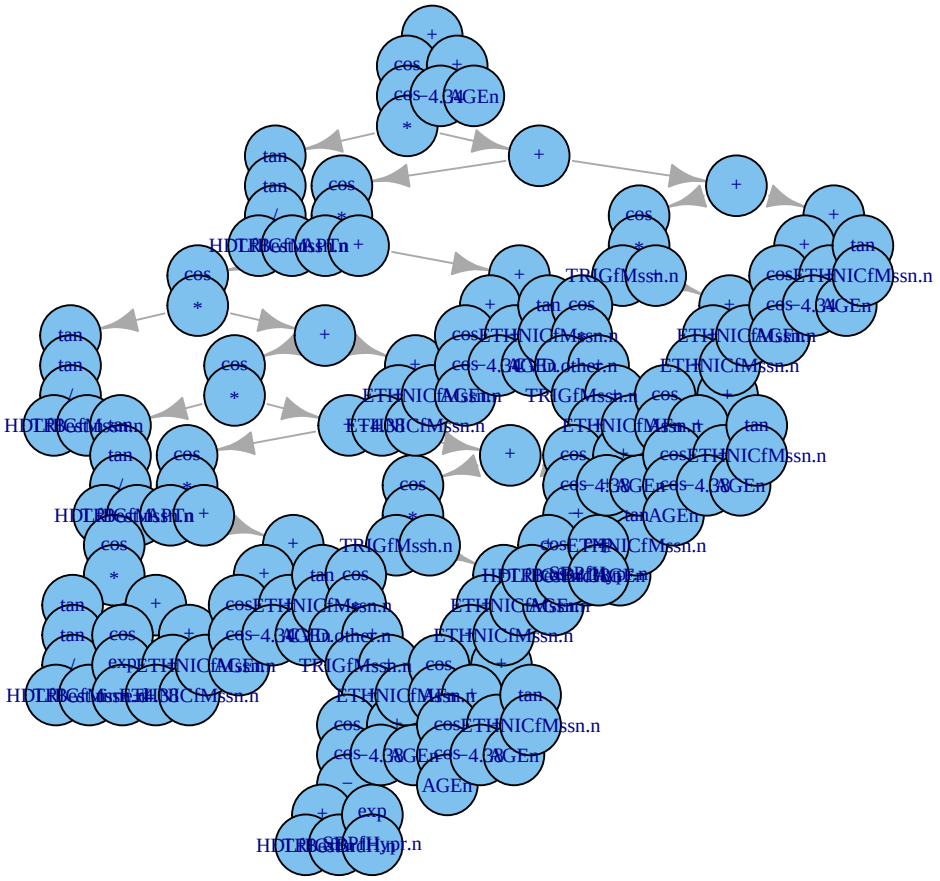

–
–
–
time.d
exp
time.d
−1.49
cos
exp
–
–
cos
AGEn
ETHNICfM.ssn.n
cos
AGEn

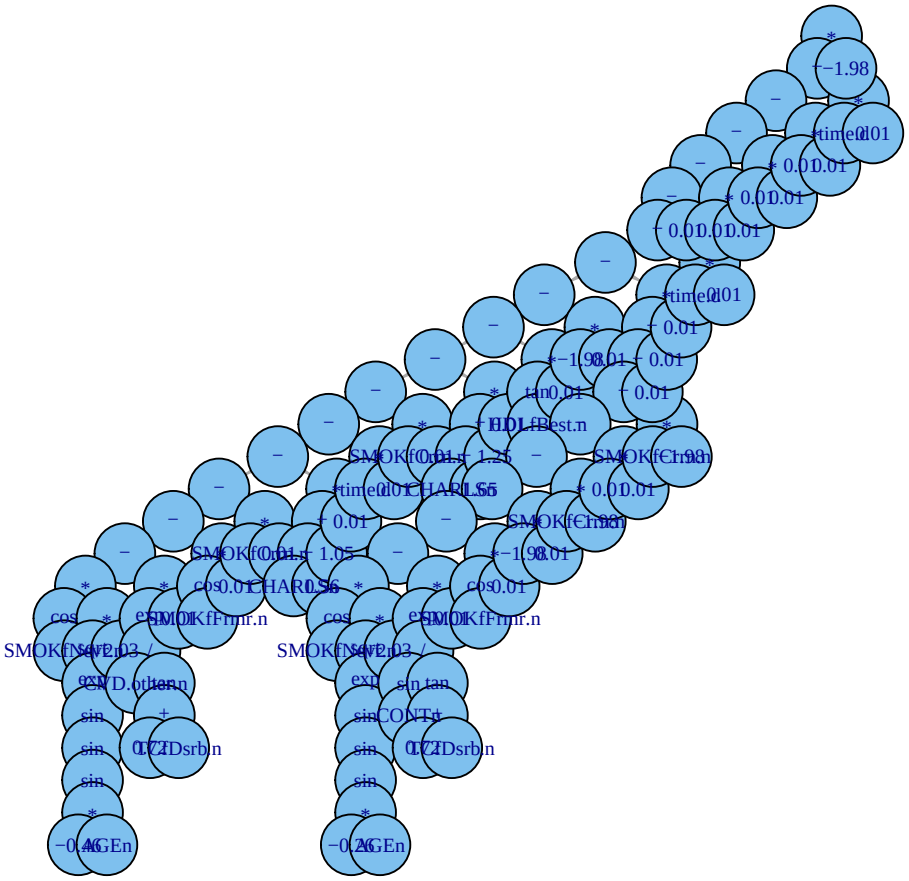

−
−2.17
/
exp
sin
sin
exp
−
/
ETHNICfMssn.n
sin
/
tan
sin
time.d
−
1.63
AGEn
tan
/
CVD.other.n
APTn
/
TRIGfMssn.n
TRIGfHigh.n

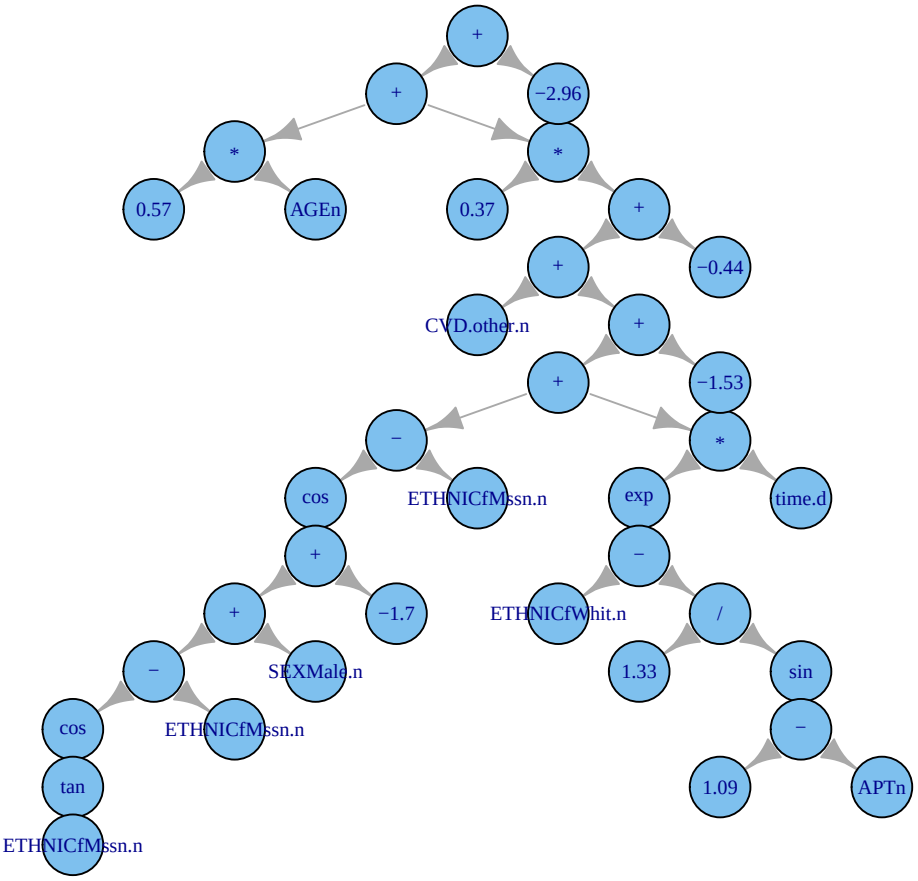

*
−
0.29
−
*
−
ETHNICfMssn.n
−2.11
*
−
ETHNICfMssn.n
cos
AGEn
−
ETHNICfMssn.n
LLTn
−
ETHNICfMssn.n
log
*
exp
exp
−
0.31
cos
−0.07
log
+
SMOKfNevr.n
exp
1.4
time.d
APTn
cos
/
CVD.other.n
LDLfMssn.n

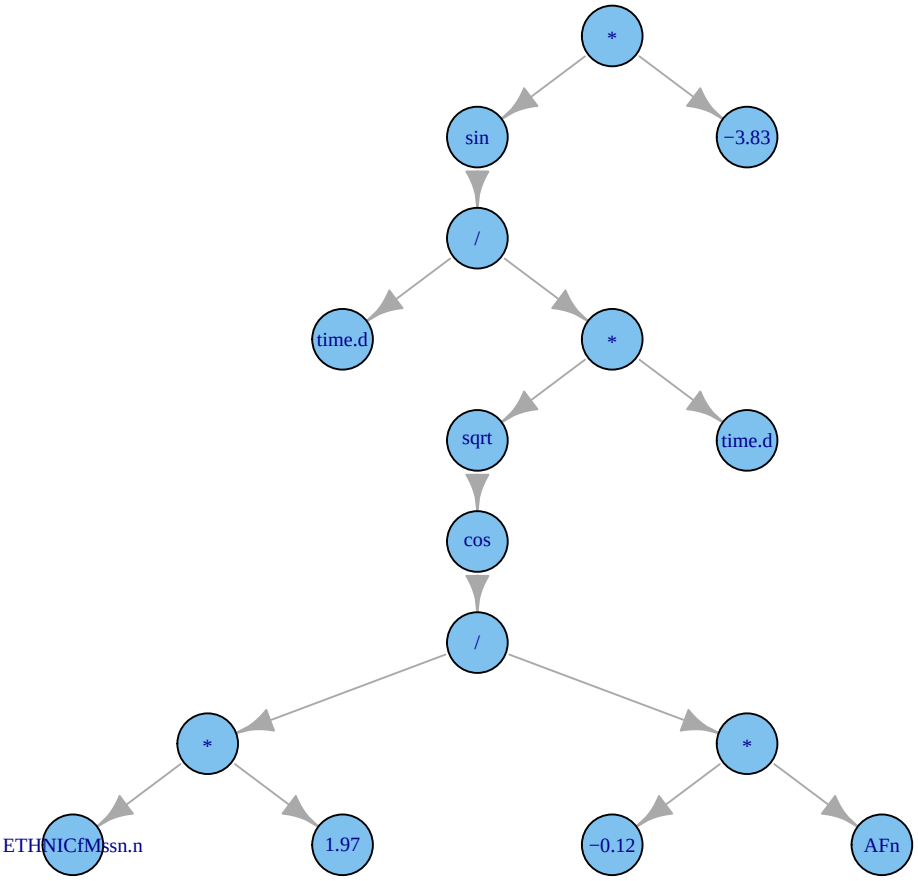

−
−3.5
−
*
sin
−
−0.69
sin
*
tan
AGEn
+
exp
ETHNICfMssn.n
sin
tan
−1.42
sin
−
time.d
CVD.other.n
*
TRIGfDsrb.n
sin
cos
*
/
exp
LDLfHigh.n
APTn
tan
/
SMOKfCrrn.n
TRIGfDsrb.n

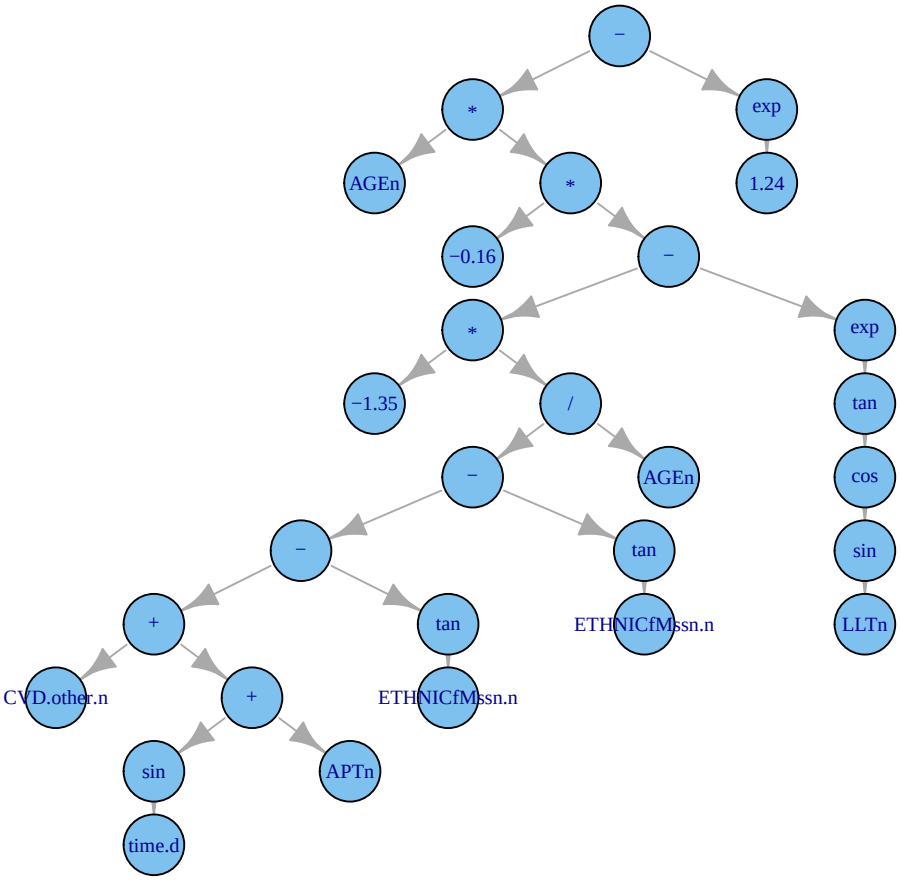

*
+
−0.89
*
3.98
+
−0.65
*
AGEn
*
exp
+
−0.89
cos
sin
sin
*
*
ETHNICfMssn.n
ETHNICfUnkn.n
/
+
sin
sqrt
CVD.other.n
sin
CVD.other.n
tan
0.38
time.d
HDLfBtl.n

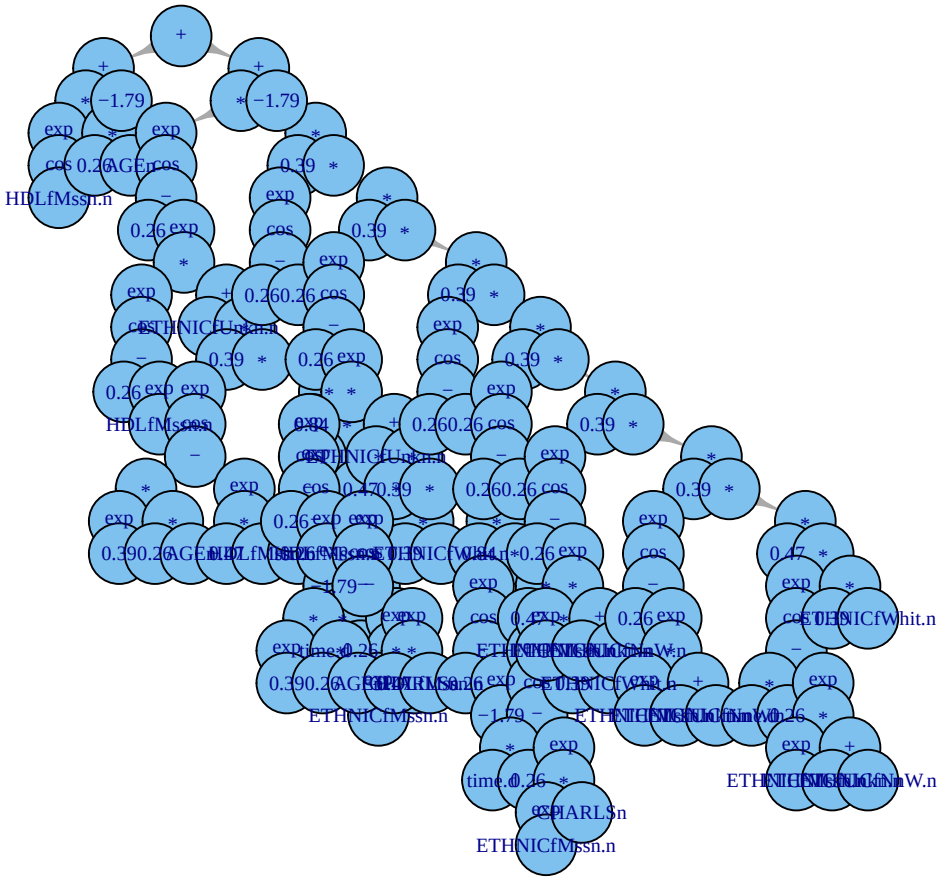


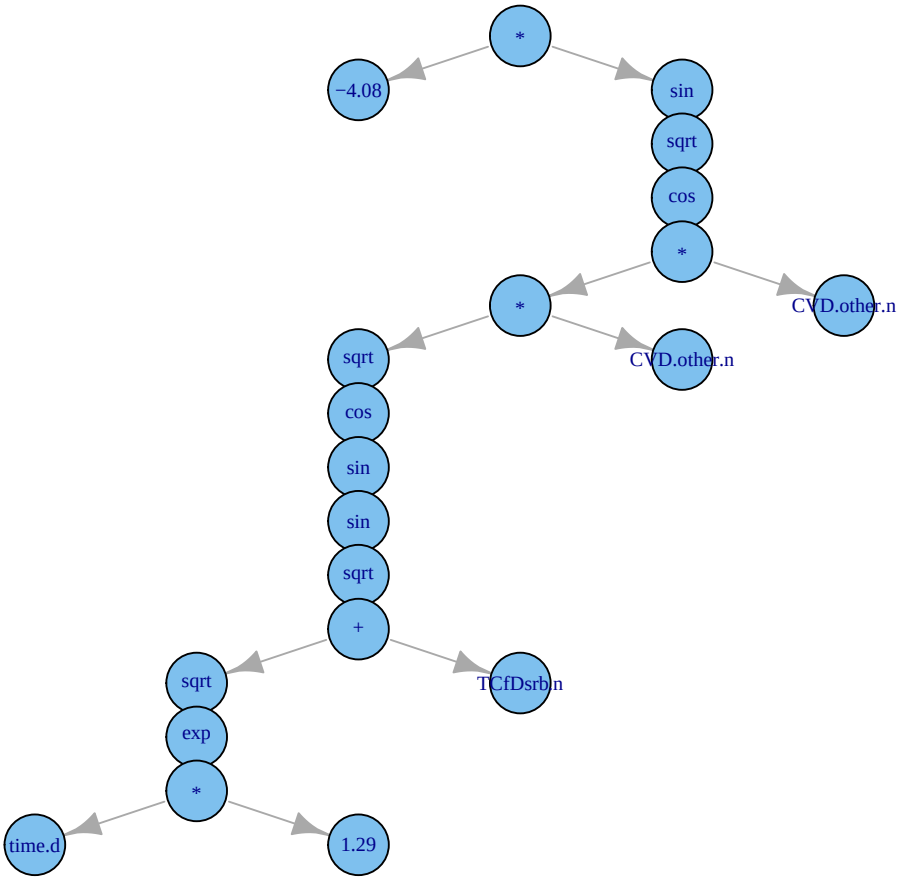



+
+
−2.96
*
*
0.57
AGEn
0.37
+
+
−0.44
CVD.other.n
+
+
−1.53
−
*
cos
ETHNICfMssn.n
exp
time.d
+
−
+
−1.7
ETHNICfWhit.n
/
−
SEXMale.n
1.33
sin
cos
ETHNICfMssn.n
−
tan
1.09
APTn
ETHNICfMssn.n


*
sin
−3.83
/
time.d
*
sqrt
time.d
cos
/
*
*
ETHNICfM.ssn.n
1.97
−0.12
AFn


−
*
exp
AGEn
*
1.24
−0.16
−
*
exp
−1.35
/
tan
−
AGEn
cos
−
tan
sin
+
tan
ETHNICfM.ssn.n
LLTn
CVD.other.n
+
ETHNICfM.ssn.n
sin
APTn
time.d


+
+
+
*
−1.79
exp
cos
0.26
AGEr
HDLfMssn.n
0.39
ETHNICfUnkn.n
ETHNICfMssn.n
ETHNICfWhit.n
ARLSn
time
0.47


*
−4.08
sin
sqrt
cos
*
*
CVD.other.n
sqrt
CVD.other.n
cos
sin
sin
sqrt
+
sqrt
TCfDsrb.n
exp
*
time.d
1.29

Appendix I

Predictor Effects: CPRD experiments

Plots of the effects of predictor values on log hazard in the 'final' GP model in the CPRD experiments in chapter 7

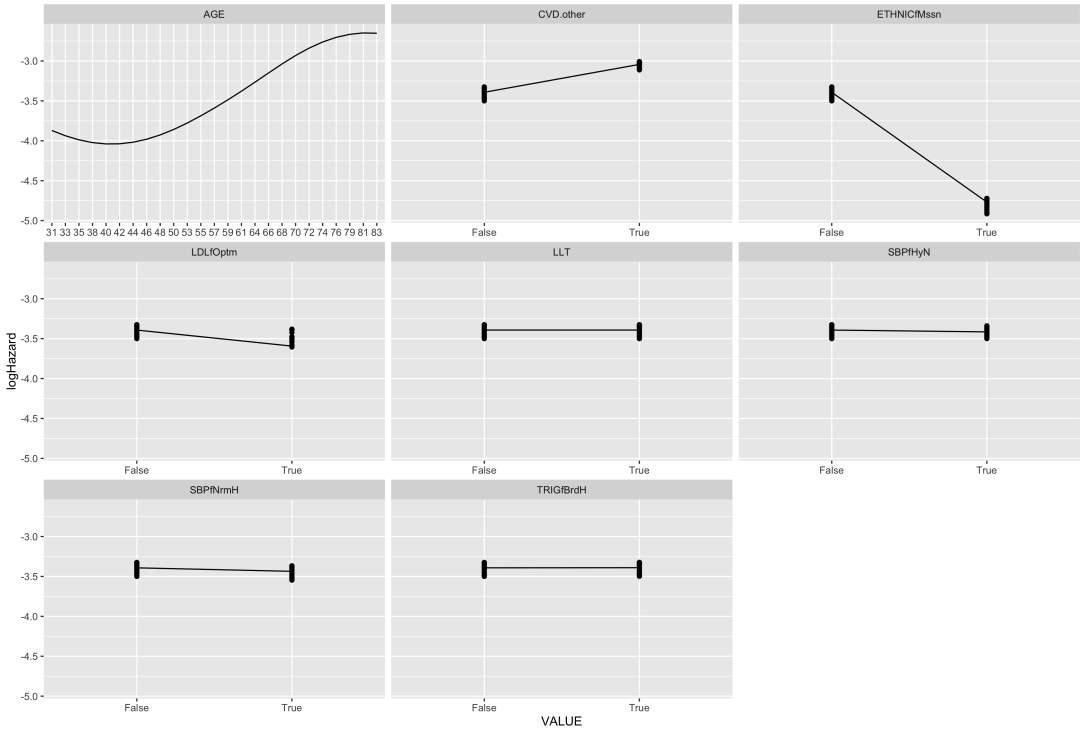


Figure I.1: Plots of the effects of predictor values on log hazard in the 'final' GP model in the CPRD experiments in chapter 7.

Appendix J

Results: CPRD experiments (secondary analysis)

Results of the additional experiments that repeat the CPRD experiments in chapter 7, but only on the subset of covariates that were selected with a relatively high frequency(> 0.5) in the main experiment.

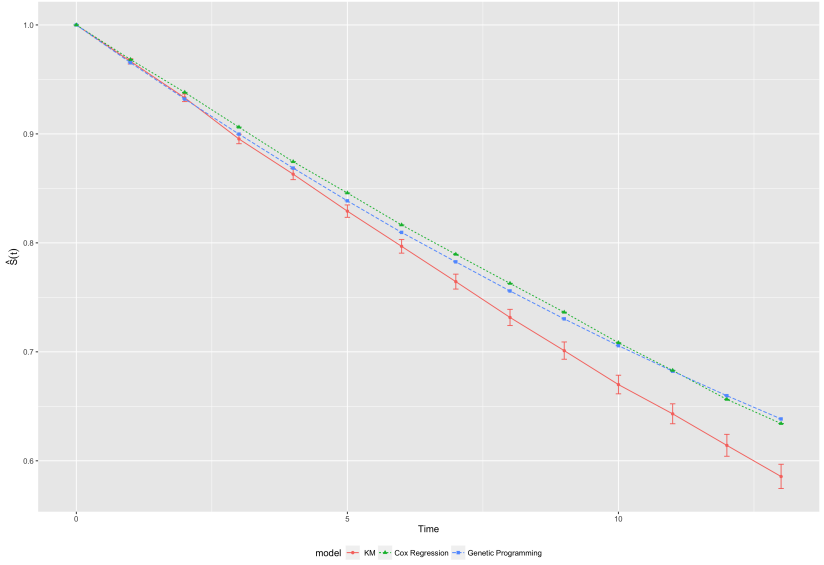


Figure J.1: Average survival curves for the Cox regression and genetic programming models. The error bars represent ± 2 standard errors of the KM estimates.

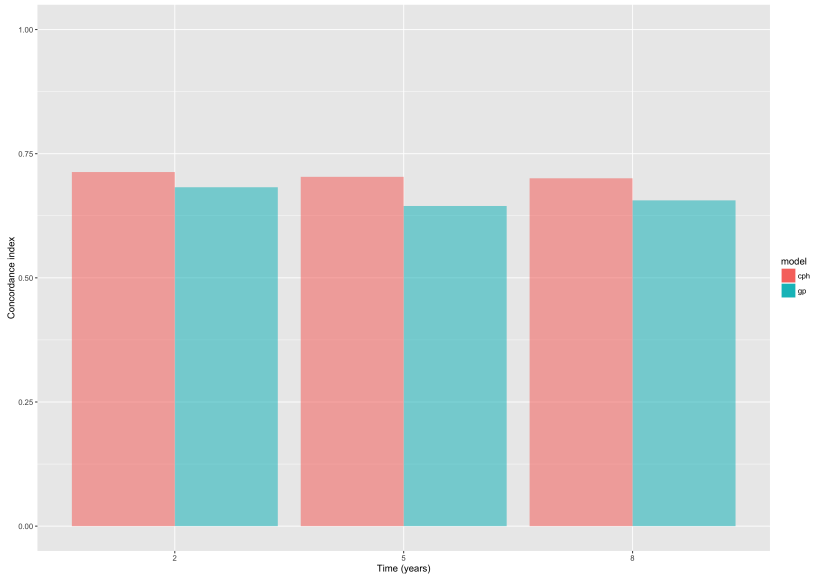


Figure J.2: C-statistic estimates by model for t=1, 3 and 5 years

Table J.1: C-statistic estimates by model at t=2, 5, and 8 years

Time (years)	Genetic Programming (superset)	Genetic Programming (subset)
2	0.656	0.682
5	0.621	0.645
8	0.631	0.656

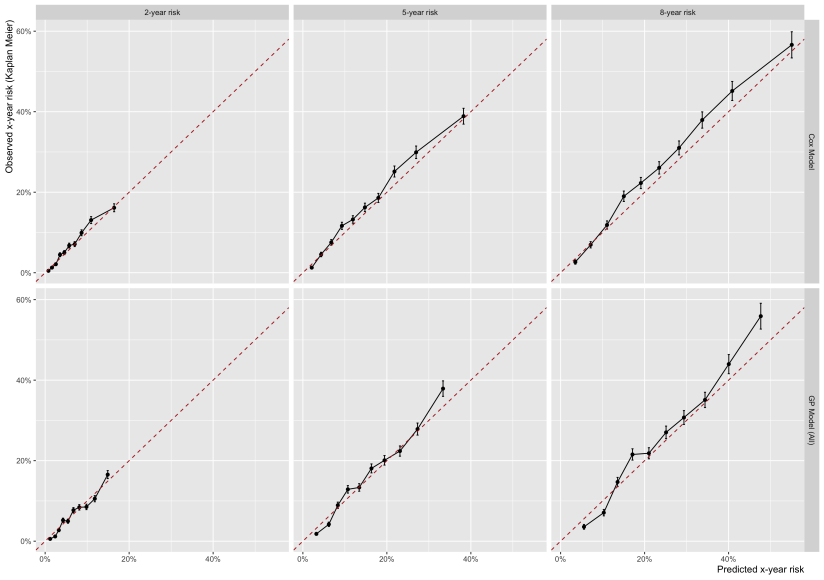


Figure J.3: Calibration plots for the Cox regression and genetic programming models, at t=1, 3, and 5 years..

Table J.2: χ^2 statistic for the comparison between observed versus expected (according to the model) number of events in groups of patients defined according to the predicted $1 - S(t)$ at t=2, 5, and 8 years.

Time (years)	Genetic Programming (superset)		Genetic Programming (subset)	
t	χ^2	p-value	χ^2	p-value
2	1589	< 0.001	1590	< 0.001
5	4146	< 0.001	4256	< 0.001
8	6937	< 0.001	7129	< 0.001

Bibliography

[1] Multiobjective Genetic Programming: Reducing Bloat Using SPEA2.

[2] AI Adler. UKPDS modelling of cardiovascular risk assessment and lifetime simulation of outcomes. *Diabet. Med*, Suppl 2:41–6, aug 2008.

[3] Ralph B D Agostino, Ramachandran S Vasan, Michael J Pencina, A Philip, Mark Cobain, Joseph M Massaro, and William B Kannel. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, 117(6):743–53, 2008.

[4] Ricardo Aler, Daniel Borrajo, and Pedro Isasi. Using genetic programming to learn and improve control knowledge. *Artif. Intell.*, 141(1-2):29–56, 2002.

[5] Jess Allen, Hazel M Davey, David Broadhurst, Jim K Heald, Jem J Rowland, Stephen G Oliver, and Douglas B Kell. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.*, 21(6):692–6, jun 2003.

[6] Paul D. Allison. *Survival Analysis using SAS: A Practical Guide*. SAS Institute, 2nd edition, 2010.

[7] L Alonso and R Schott. *Random Generation of Trees*. Kluwer Academic Publishers, 1995.

[8] K. M. Anderson, P. W. Wilson, P. M. Odell, and W. B. Kannel. An updated coronary risk profile. A statement for health professionals. *Circulation*, 83(1):356–362, jan 1991.

[9] Keaven M Anderson, Patricia M Odell, Peter W F Wilson, William B Kannel, and M P H Framingham. Cardiovascular disease risk profiles. *Am. Heart J.*, 121(1 Pt 2):293–8, jan 1991.

[10] Peter J Angeline. An investigation into the sensitivity of genetic programming to the frequency of leaf selection during subtree crossover. In *GECCO '96 Proc. First Annu. Conf. Genet. Program.*, pages 21–29, 1996.

[11] G. Assmann. Simple Scoring Scheme for Calculating the Risk of Acute Coronary Events Based on the 10-Year Follow-Up of the Prospective Cardiovascular Munster (PROCAM) Study. *Circulation*, 105(3):310–315, jan 2002.

[12] G Assmann, H Schulte, P Cullen, and U Seedorf. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Münster (PROCAM) study. *Eur. J. Clin. Invest.*, 37(12):925–932, 2007.

[13] A Bagust, PK Hopkinson, L Maslove, and CJ Currie. The projected health care burden of Type 2 diabetes in the UK from 2000 to 2060. *Diabet Med.*, 19(Suppl 4):1–5, jul 2002.

[14] Christian A Bannister, Craig J Currie, Alun Preece, and Irena Spasic. Automatic development of clinical prediction models with genetic programming: a case study in cardiovascular disease. In *Value Heal.*, pages 17:A200–1, 2014.

[15] Christian A Bannister, Chris D Poole, Sara Jenkins-Jones, Christopher Ll Morgan, Glyn Elwyn, Irena Spasic, and Craig J Currie. External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations. *Diabetes Care*, 37(2):537–45, feb 2014.

[16] Christian A Bannister, Christopher D Poole, Sara Jenkins-Jones, Christopher Ll Morgan, Glyn Elwyn, and Craig J Currie. Validation of UKPDS Risk Engine Predictions Among Patients with Type 2 Diabetes Routinely Managed in UK Primary Care. In *Diabetes*, pages 62(suppl 1):A1–A98 276–OR, 2013.

[17] Wolfgang Banzhaf, Peter Nordin, Robert Keller, and Frank Francone. *Genetic Programming: An Introduction On the Automatic Evolution of Computer Programs and Its Application*. 1998.

[18] S J Barrett. Predicting Biochemical Interactions: Human P450 2D6 Enzyme Inhibition. (Section 6):8–12, 2003.

[19] S. J. Barrett. Recurring Analytical Problems within Drug Discovery and Development. In Tobias Scheffer and Ulf Leser, editors, *Data Min. Text Min. Bioinforma. Proc. Eur. Work.*, pages 6–7, 2003.

[20] Nicola Beume, Boris Naujoks, and Michael Emmerich. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *Eur. J. Oper. Res.*, 181(3):1653–1669, 2007.

[21] Authur S Bickel and Riva Wenig Bickel. Tree Structured Rules in Genetic Algorithms. In *Genet. Algorithms their Appl. Proc. Second Int. Conf. Genet. Algorithms*, pages 77–81, 1987.

[22] Cornelis J Biesheuvel, Ivar Siccama, Diederick E Grobbee, and Karel G M Moons. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. *J. Clin. Epidemiol.*, 57(6):551–60, jun 2004.

[23] E Biganzoli, P Boracchi, L Mariani, and E Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat. Med.*, 17(10):1169–1186, 1998.

[24] Walter Bohm and Andreas Geyer-Schulz. Exact Uniform Initialization for Genetic Programming. In *Found. Genet. Algorithms IV*, pages 379–409, 1996.

[25] C C Bojarczuk, H S Lopes, and A A Freitas. Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Eng. Med. Biol. Mag.*, 19(4):38–44, 2000.

[26] J Boye, L Geiss, and A Honeycutt. Projection of Diabetes Burden Through 2050. *Diabetes Care*, 24(11), 2001.

[27] Markus F. Brameier and Wolfgang Banzhaf. *Linear Genetic Programming*. Springer US, 1 edition, 2007.

[28] British Medical Association and Royal Pharmaceutical Society. British National Formulary , 2012.

[29] Jason Brownlee. Clever Algorithms: Nature-Inspired Programming Recipes. 2011.

[30] Weihua Cai, Arturo Pacheco-Vega, Mihir Sen, and K.T. Yang. Heat transfer correlations by symbolic regression. *Int. J. Heat Mass Transf.*, 49(23-24):4352–4359, nov 2006.

[31] Canadian Diabetes Association. The prevalence and costs of diabetes.

[32] Canadian Diabetes Association. Clinical Practice Guidelines for the Prevention and Management of Diabetes in Canada. *Can J Diabetes*, 32(Supplement 1), 2008.

[33] Centers for Disease Control and Prevention. National Diabetes Fact Sheet 2011, 2011.

[34] Lloyd E Chambless, Aaron R Folsom, A.Richey Sharrett, Paul Sorlie, David Couper, Moyses Szklo, and F.Javier Nieto. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. *J. Clin. Epidemiol.*, 56(9):880–890, sep 2003.

[35] P Chamnan, R K K Simmons, S J J Sharp, S J J Griffin, and N J J Wareham. Cardiovascular risk assessment scores for people with diabetes: a systematic review. *Diabetologia*, 52(10):2001–14, oct 2009.

[36] C Chateld. Model uncertainty, data mining and statistical inference," JR Statist. *J R Stat. Soc*, 158(3):419–466, 1995.

[37] Kumar Chellapilla. Evolving computer programs without subtree crossover. *IEEE Trans. Evol. Comput.*, 1(3):209–216, 1997.

[38] Gabriel Chodick, Anthony D Heymann, Francis Wood, and Ehud Kokia. The direct medical cost of diabetes in Israel. *Eur. J. Health Econ.*, 6(2):166–71, jun 2005.

[39] Clinical Practice Research Datalink (CPRD). Clinical Practice Research Datalink, 2013.

[40] Carlos A. Coello Coello. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: A survey of the state of the art. *Comput. Methods Appl. Mech. Eng.*, 191(11-12):1245–1287, 2002.

[41] Carlos A. Coello Coello and Efrén Mezura Montes. Constraint-handling in genetic algorithms through the use of dominance-based tournament selection. *Adv. Eng. Informatics*, 16(3):193–203, 2002.

[42] Carlos Artemio Coello Coello. A Survey of Constraint Handling Techniques used with Evolutionary Algorithms. *Lania-RI-99-04, Lab. Nac. ...*, pages 1–33, 1999.

[43] David W. Coit and Alice E. Smith. Penalty Guided Genetic Search For Reliability Design Optimization. *Comput. Ind. Eng.*, 30:895–904, 1996.

[44] Pierre Collet. Genetic programming. In *Handb. Res. Nature-Inspired Comput. Econ. Manag.*, pages 59–73. Idea Group Inc, 2007.

[45] David Collett. *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science, 2nd edition, 2003.

[46] R M Conroy, K Pyörälä, A P Fitzgerald, S Sans, A Menotti, G De Backer, D De Bacquer, P Ducimetière, P Jousilahti, U Keil, I Njølstad, R G Oganov, T Thomsen, H Tunstall-Pedoe, A Tverdal, H Wedel, P Whincup, L Wilhelmsen, and I M Graham. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur. Heart J.*, 24(11):987–1003, jun 2003.

[47] A Coulter. Partnerships with patients: the pros and cons of shared clinical decision-making. *J. Health Serv. Res. Policy*, 2(2):112–21, apr 1997.

[48] DR Cox. Regression Models and Life-Tables. *J. R. Stat. Soc.*, 34(2):187–220, 1972.

[49] B.G.W. Craenen, A.E. Eiben, and E. Marchiori. How to Handle Constraints with Evolutionary Algorithms. *Pract. Handb. Genet. Algorithms Appl.*, pages 341–362, 2001.

[50] N. L. Cramer. A representation for the adaptive generation of simple sequential programs. In J. J. Grefenstette, editor, *1st Int. Conf. Genet. Algorithms*, pages 183–187, 1985.

[51] Nl Cramer. A representation for the adaptive generation of simple sequential programs. *Proc. First Int. Conf. ...*, pages 1–6, 1985.

[52] Crisis. Policy Brief: Critical Condition. Technical report, 2002.

[53] C J Currie, E a M Gale, and C D Poole. Estimation of primary care treatment costs and treatment efficacy for people with Type 1 and Type 2 diabetes in the United Kingdom from 1997 to 2007*. *Diabet Med*, 27(8):938–48, aug 2010.

[54] Craig J Currie, Chris D Poole, Marc Evans, John R Peters, and Christopher Ll Morgan. Mortality and other important diabetes-related outcomes with insulin vs other antihyperglycemic therapies in type 2 diabetes. *J Clin Endocrinol Metab*, 98(2):668–77, feb 2013.

[55] Natasha Curry, J. Billings, B. Darin, J. Dixon, M. Williams, and D. Wennberg. Predictive risk project: literature review. *New York*, (June):1–24, 2005.

[56] R.B. D'Agostino and B.H. Nam. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In N. Balakrishnan and C. Rao, editors, *Handb. Stat.*, pages 1–26. Elsevier, Amsterdam, 23 edition, 2004.

[57] C Darwin. On the origin of species by means of natural selection, or the preservation of favoured races in . . . , 1859.

[58] Wendy a Davis, Stephen Colagiuri, and Timothy M E Davis. Comparison of the Framingham and United Kingdom Prospective Diabetes Study cardiovascular risk equations in Australian patients with type 2 diabetes from the Fremantle Diabetes Study. *Med J Aust*, 190(4):180–4, feb 2009.

[59] Jennifer P Day, Douglas B Kell, and Gareth W Griffith. Differentiation of phytophthora infestans sporangia from other airborne biological particles by flow cytometry. *Appl. Environ. Microbiol.*, 68(1):37–45, 2002.

[60] E. D. de Jong, R. A. Watson, and J. B. Pollack. Reducing bloat and promoting diversity using multi-objective methods. In *Proc. Genet. Evol. Comput. Conf.*, pages 11–18, 2001.

[61] Edwin D. De Jong and Jordan B. Pollack. Multi-objective methods for tree size control. *Genet. Program. Evolvable Mach.*, 4(3):211–233, 2003.

[62] Janaína S. de Sousa, Lalinka de C. T. Gomes, George B. Bezerra, Leandro N. de Castro, and Fernando J. Von Zuben. An Immune-Evolutionary Algorithm for Multiple Rearrangements of Gene Expression Data. *Genet. Program. Evolvable Mach.*, 5(2):157–179, jun 2004.

[63] Postdischarge Death, Kim a Eagle, Michael J Lim, Omar H Dabbous, Karen S Pieper, Robert J Goldberg, Shaun G Goodman, Christopher B Granger, P Gab-

riel Steg, Joel M Gore, Marcus D Flather, and Keith a a Fox. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA*, 291(22):2727–2733, 2004.

[64] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197, 2002.

[65] Ian Dempsey, Michael O'Neill, and Anthony Brabazon. *Foundations in Grammatical Evolution for Dynamic Environments*. Springer, 2009.

[66] Department of Health. Putting prevention first - vascular checks: risk assessment and management. *London Dep. Heal.*, 2008.

[67] Department of Health. Department of Health Spending Review 2010, 2010.

[68] Department of Health. Hospital Episode Statistics, 2012.

[69] P D'haeseleer. Context preserving crossover in genetic programming. In *Proc. 1994 IEEE World Congr. Comput. Intell.*, pages 256—-261, 1994.

[70] Diabetes UK. Diabetes in the UK 2012: Key statistics on diabetes 2012, 2012.

[71] RF Dijkstra and LW Niessen. Patient centred and professional directed implementation strategies for diabetes guidelines: a cluster randomized trial based cost effectiveness analysis. *Diabet Med*, 23(2):164–70, feb 2006.

[72] Peter T Donnan, Louise Donnelly, John P New, and Andrew D Morris. Derivation and validation of a prediction score for major coronary heart disease events in a U.K. type 2 diabetic population. *Diabetes Care*, 29(6):1231–6, jun 2006.

[73] D Draper. Assessment and Propagation of Model Uncertainty. *J R Stat. Soc*, 57(1):45–97, 1995.

[74] B Efron. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc*, 72(359):557–565, 1977.

[75] B Efron and R Tibshirani. *An introduction to the bootstrap. Monographs on statistics and applied probability*. New York: Chapman & Hall, 1993.

[76] J Eggermont. *Data Mining using Genetic Programming: Classification and Symbolic Regression*. PhD thesis, 2005.

[77] Aniko Ekart and S Z Nemeth. Selection Based on the Pareto Nondomination Criterion for Controlling Code Growth in Genetic Programming. *Genet. Program. Evolvable Mach.*, 2(1):61–73, 2001.

[78] C Raina Elley, Elizabeth Robinson, Tim Kenealy, Dale Bramley, and Paul L Drury. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the new zealand diabetes cohort study. *Diabetes Care*, 33(6):1347–52, jun 2010.

[79] David I. Ellis, David Broadhurst, Douglas B. Kell, Jem J. Rowland, and Royston Goodacre. Rapid and Quantitative Detection of the Microbial Spoilage of Meat by Fourier Transform Infrared Spectroscopy and Machine Learning. *Appl. Environ. Microbiol.*, 68(6):2822, 2002.

[80] R Eriksson and B Olsson. Adapting genetic regulatory models by genetic programming. *Biosystems.*, 76(1-3):217–27, 2004.

[81] Chris Feudtner, Kari R Hexem, Mayadah Shabbout, James a Feinstein, Julie Sochalski, and Jeffery H Silber. Prediction of pediatric death in the year after hospitalization: a population-level retrospective cohort study. *J. Palliat. Med.*, 12(2):160–169, 2009.

[82] First Databank. Multilex Drug Knowledge Base , 2012.

[83] Oliver Flasch. *A Modular Genetic Programming System*. PhD thesis, 2015.

[84] Oliver Flasch, Olaf Mersmann, Thomas Bartz-Beielstein, Joerg Stork, and Martin Zaefferer. rgp: R genetic programming framework, 2014.

[85] D Fogel. Evolving computer programs. In D Fogel, editor, *Evol. Comput. Foss. Rec.*, chapter 5, pages 143–144. MIT Press, 1998.

[86] L. Fogel, A. Owens, and M Walsh. *Artificial Intelligence through Simulated Evolution.* John Wiley, 1966.

[87] L. J. Fogel. Autonomous automata. *Ind. Res.*, 4:14–19, 1962.

[88] L. J. Fogel. *On the Organization of Intellect*. Phd thesis, UCLA, 1964.

[89] S Forrest. Genetic algorithms: principles of natural selection applied to computation. *Science (80-.).*, 261(5123):872–878, 1993.

[90] Alex Alves Freitas. Evolutionary Computation. In W. Klosgen and J. Zytkow, editors, *Handb. Data Min. Knowl. Discov.*, volume 4, pages 698–706. Oxford University Press, 2002.

[91] R M Friedberg. A Learning Machine : Part I. (January), 1958.

[92] van Staa TP Gallagher AM, Puri S. Linkage of the General Practice Research Database (GPRD) with Other Data Sources. *Pharmacoepidemiol Drug Saf*, 20:S230, 2011.

[93] Thomas a Gaziano, Cynthia R Young, Garrett Fitzmaurice, Sidney Atwood, and J Michael Gaziano. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. *Lancet*, 371(9616):923–31, mar 2008.

[94] R J Gilbert, R Goodacre, a M Woodward, and D B Kell. Genetic programming: a novel method for the quantitative analysis of pyrolysis mass spectral data. *Anal. Chem.*, 69(21):4381–9, nov 1997.

[95] C Glümer, M Yuyun, S Griffin, D Farewell, D Spiegelhalter, a L Kinmonth, and N J Wareham. What determines the cost-effectiveness of diabetes screening? *Diabetologia*, 49(7):1536–44, jul 2006.

[96] David Edward Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, 1989.

[97] R Goodacre, B Shann, R J Gilbert, E M Timmins, a C McGovern, B K Alsberg, D B Kell, and N a Logan. Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Anal. Chem.*, 72(1):119–27, jan 2000.

[98] Royston Goodacre. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vib. Spectrosc.*, 32(1):33–45, aug 2003.

[99] Christopher B Granger, Robert J Goldberg, Omar Dabbous, Karen S Pieper, Kim a Eagle, Christopher P Cannon, Frans Van De Werf, Alvaro Avezum, Shaun G Goodman, Marcus D Flather, and Keith a a Fox. Predictors of hospital mortality in the global registry of acute coronary events. *Arch. Intern. Med.*, 163(19):2345–2353, 2003.

[100] Jeremy Gray, Douglas Orr, and Azeem Majeed. Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers. *BMJ*, 326(7399):1130, 2003.

[101] S. Gustafson, E.K. Burke, and N. Krasnogor. On improving genetic programming for symbolic regression. In *Evol. Comput. 2005. 2005 IEEE Congr.*, volume 1, pages 912–919, 2005.

[102] GH Guyatt, RB Haynes, and RZ Jaeschke. Users' Guides to the Medical LiteratureXXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care. *Jama*, 284(10):1290–1296, 2000.

[103] R N Guzder, W Gatling, M a Mullee, R L Mehta, and C D Byrne. Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed Type 2 diabetes: results from a United Kingdom study. *Diabet Med*, 22(5):554–62, may 2005.

[104] W Hamilton, R Lancashire, D Sharp, T J Peters, K K Cheng, and T Marshall. The importance of anaemia in diagnosing colorectal cancer: a case-control study using electronic primary care records. *Br. J. Cancer*, 98(2):323–7, jan 2008.

[105] Tarek A Hammad, Mara A McAdams, Andrea Feight, Solomon Iyasu, and Gerald J Dal Pan. Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. *Pharmacoepidemiol Drug Saf*, 17(12):1197–201, dec 2008.

[106] Simon Handley. Automatic Learning of a Detector for alpha-helices in Protein Sequences Via Genetic Programming. In Stephanie Forrest, editor, *Proc. 5th Int. Conf. Genet. Algorithms, ICGA-93*, pages 271–278. Morgan Kaufmann, 1993.

[107] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[108] Rachael Harker. NHS Expenditure - Commons Library Standard Note. Technical report, 2011.

[109] FE Harrell, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[110] FE Harrell, KL Lee, RM Califf, DB Pryor, and RA Rosati. Regression modelling strategies for improved prognostic prediction. *Stat Med*, 3(2):143–52, 1984.

[111] FE Harrell, KL Lee, and DB Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, 15:361–387, 1996.

[112] Frank E Harrell. *REGRESSION MODELING STRATEGIES with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York, second ed edition, 2006.

[113] Kim Harries and Peter Smith. Exploring Alternative Operators and Search Strategies in Genetic Programming. In *Proc. Genet. Program.*, pages 147–155, 1997.

[114] George G Harrigan, Roxanne H LaPlante, Greg N Cosma, Gary Cockerell, Royston Goodacre, Jane F Maddox, James P Luyendyk, Patricia E Ganey, and Robert a Roth. Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicol. Lett.*, 146(3):197–205, feb 2004.

[115] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer; 2nd ed. 2009. Corr. 3rd printing 5th Printing. edition (20 April 2011), 2009.

[116] Health and Social Care Information Centre. Quality and Outcomes Framework , Achievement , prevalence and exceptions data, 2011/12. Technical report, 2012.

[117] a Geert Heidema, Jolanda M a Boer, Nico Nagelkerke, Edwin C M Mariman, Daphne L van der A, and Edith J M Feskens. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet.*, 7:23, jan 2006.

[118] Emily Herrett, Sara L Thomas, W Marieke Schoonen, Liam Smeeth, and Andrew J Hall. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol*, 69(1):4–14, jan 2010.

[119] Mark Hinchliffe, Mark Willis, and Ming Tham. Chemical Process Sytems Modelling Using Multi-objective Genetic Programming. In *Genet. Program. 1998 Proc. Third Annu. Conf.*, pages 134–139, 1998.

[120] J. Hippisley-Cox, C. Coupland, J. Robson, a. Sheikh, and P. Brindle. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *Bmj*, 338(mar17 2):b880–b880, mar 2009.

[121] J Hippisley-Cox, C Coupland, Y Vinogradova, J Robson, and P Brindle. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart*, 94(1):34–9, jan 2008.

[122] Julia Hippisley-Cox, Carol Coupland, John Robson, and Peter Brindle. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Br. Med. J.*, 341(dec09 1):c6624, 2010.

[123] Julia Hippisley-Cox, Carol Coupland, John Robson, and Peter Brindle. Advantages of QRISK2 (2010): the key issue is ethnicity and extent of reallocation. *Heart*, 97(6):515; author reply 515–6, mar 2011.

[124] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, 335(7611):136, 2007.

[125] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Rubin Minhas, Aziz Sheikh, and Peter Brindle. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336(7659):1475–1482, 2008.

[126] Shinn-Ying Ho, Chih-Hung Hsieh, Hung-Ming Chen, and Hui-Ling Huang. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems.*, 85(3):165–76, sep 2006.

[127] Sarah H Holden, Anthony H Barnett, John R Peters, Sara Jenkins-Jones, Chris D Poole, Christopher Ll Morgan, and Craig J Currie. The incidence of type 2 diabetes in the United Kingdom from 1991 to 2010. *Diabetes Obes Metab*, (in press, 2013.

[128] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[129] Paul Holmes. The Odin Genetic Programming System. Technical Report RR-95-3, 1995.

[130] Jin-Hyuk Hong and Sung-Bae Cho. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. *Artif. Intell. Med.*, 36(1):43–58, jan 2006.

[131] David W. Hosmer Jr and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, 2 edition, 2000.

[132] Hitoshi Iba. Random Tree Generation for Genetic Programming. In *Parallel Probl. Solving from Nat. IV, Proc. Int. Conf. Evol. Comput.*, pages 144–153, 1996.

[133] In Practice Systems. Vision, 2012.

[134] International Diabetes Federation. IDF Clinical Guidelines Task Force. Global Guideline for Type 2 Diabetes: recommendations for standard, comprehensive, and minimal care. *Diabet Med*, 23:579–593, 2006.

[135] International Health Terminology Standard Development Organisation. SNOMED CT, 2012.

[136] Kristel J M Janssen, Ivar Siccama, Yvonne Vergouwe, Hendrik Koffijberg, T P a Debray, Maarten Keijzer, Diederick E Grobbee, and Karel G M Moons. Development and validation of clinical prediction models: marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming. *J. Clin. Epidemiol.*, 65(4):404–12, apr 2012.

[137] Kristel J M Janssen, Yvonne Vergouwe, a Rogier T Donders, Frank E Harrell, Qingxia Chen, Diederick E Grobbee, and Karel G M Moons. Dealing with missing predictor values when applying clinical prediction models. *Clin. Chem*, 55(5):994–1001, may 2009.

[138] J M Jerez, L Franco, E Alba, a Llombart-Cussac, a Lluch, N Ribelles, B Munárriz, and M Martín. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Res. Treat.*, 94(3):265–72, dec 2005.

[139] J. Jimeno Mollet, N. Molist Brunet, J. Franch Nadal, V. Serrano Borraz, L. Serrano Barragán, and R. Gracia Giménez. Variabilidad en la estimación del riesgo coronario en la diabetes mellitus tipo 2 [Variability in the calculation of coronary risk in type-2 diabetes mellitus]. *Aten Primaria*, 35(1):30–36, jan 2005.

[140] H. E. Johnson, R. J. Gilbert, M. K. Winson, R. Goodacre, A. R. Smith, J. J. Rowland, M. A. Hall, and D. B. Kell. Explanatory Analysis of the Metabolome Using Genetic Programming of Simple, Interpretable Rules. *Genet. Program. Evolvable Mach.*, 1:243–258, 2000.

[141] a Jones, D Young, J Taylor, D B Kell, and J J Rowland. Quantification of microbial productivity via multi-angle light scattering and supervised learning. *Biotechnol. Bioeng.*, 59(2):131–43, jul 1998.

[142] B Jönsson. Revealing the cost of Type II diabetes in Europe. *Diabetologia*, 45(7):S5–12, jul 2002.

[143] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, 2nd edition, 2002.

[144] WB B Kannel, D McGee, and T Gordon. A general cardiovascular risk profile: the Framingham Study. *Am. J. Cardiol.*, 36(July):46–51, 1976.

[145] Wolfgang Kantschik and Wolfgang Banzhaf. Linear-Tree GP and its comparison with other GP structures. *Genet. Program. Proc. EuroGP2001*, 2038:302–312, 2001.

[146] Wolfgang Kantschik and Wolfgang Banzhaf. Linear-Graph GP: A New GP Structure. *Genet. Program. Proc. 5th Eur. Conf. EuroGP 2002*, 2278:83–92, 2002.

[147] EL Kaplan and P Meier. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481, 1958.

[148] Michael W Kattan. Comparison of Cox regression with other methods for determining prediction models and nomograms. *J. Urol.*, 170(6 Pt 2):S6–9; discussion S10, dec 2003.

[149] Maarten Keijzer. Scaled Symbolic Regression. *Genet. Program. Evolvable Mach.*, 5(3):259–269, sep 2004.

[150] Maarten Keijzer and James Foster. Crossover bias in genetic programming. *Genet. Program.*, 4445:33–44, 2007.

[151] D B Kell, R M Darby, and J Draper. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.*, 126(3):943–51, jul 2001.

[152] DB Kell. Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.*, 29(1-2):131–143, 2002.

[153] Douglas Kell. Defence against the flood. *Bioinforma. World*, (February):16–18, 2002.

[154] Douglas B Kell. Genotype-phenotype mapping: genes as computer programs. *Trends Genet.*, 18(11):555–9, dec 2002.

[155] AP Kengne, A Patel, S Colagiuri, and S Heller. The Framingham and UK Prospective Diabetes Study (UKPDS) risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: the Action inDiabetes and Vascular Disease: Preterax . *Diabetologia*, 53:821–831, 2010.

[156] Louise C. Kenny, Warwick B. Dunn, David I. Ellis, Jenny Myers, Philip N. Baker, and Douglas B. Kell. Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics*, 1(3):227–234, sep 2005.

[157] Nada F Khan, Sian E Harrison, and Peter W Rose. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pr.*, 60(572):e128–36, mar 2010.

[158] Dokyoon Kim, Ruowang Li, Scott M. Dudek, and Marylyn D. Ritchie. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *J. Biomed. Inform.*, 56(June):1–9, 2015.

[159] H King, R Aubert, and W German. Global Burden of Diabetes, 1995-2025. *Diabetes Care*, 21(9):1414–1431, 2000.

[160] Kenneth E. Kinnear. Evolving a sort: Lessons in genetic programming. *IEEE Int. Conf. Neural Networks - Conf. Proc.*, 1993-Janua:881–888, 1993.

[161] Kenneth E Kinnear Jr. Fitness landscapes and difficulty in genetic programming. *Evol. Comput. 1994. IEEE World Congr. Comput. Intell. Proc. First IEEE Conf.*, pages 142–147, 1994.

[162] David G. Kleinbaum and Mitchel Klein. *Survival Analysis, a Self-Learning Text*. Springer, third edition, 2011.

[163] J Klockgether and H-P Schwefel. Two-phase nozzle and hollow core jet experiments. In *Elev. Symp. Eng. Asp. Magnetohydrodyn.*, pages 141–148, California Institute of Technology, 1970.

[164] J. a. Knottnerus. Application of Logistic Regression to the Analysis of Diagnostic Data: Exact Modeling of a Probability Tree of Multiple Binary Variables. *Med. Decis. Mak.*, 12(2):93–108, jun 1992.

[165] V. Kothari. UKPDS 60: Risk of Stroke in Type 2 Diabetes Estimated by the UK Prospective Diabetes Study Risk Engine. *Stroke*, 33(7):1776–1781, jul 2002.

[166] J R Koza. Hierarchical genetic algorithms operating on populations of computer programs. *Proc. Elev. Int. Jt. Conf. Artif. Intell. IJCAI-89*, 1:768–774, 1989.

[167] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, MA, USA, 1992.

[168] J. R. Koza, L. W. Jones, M. A. Keane, and M. J. Streeter. Toward Automated Design of Industrial-Strength Analog Circuits by Means of Genetic Programming. In Ann Arbor Springer, editor, *Genet. Program. Theory Pract. II*, chapter 8. 2005.

[169] John R Koza and David Andre. Classifying Protein Segments as Transmembrane Domains Using Architecture-Altering Operations in Genetic Programming. In Peter J. Angeline and Kenneth E. Kinnear, editors, *Adv. Genet. Program. II*, chapter 8, pages 155–176. MIT Press, Cambridge, MA, USA, 1996.

[170] D.H. Kraft, F. E. Petry, W. P. Buckles, and T. Sadasivan. The use of genetic programming to build queries for information retrieval. In *Evol. Comput. 1994. IEEE World Congr. Comput. Intell. Proc. First IEEE Conf.*, volume 1, pages 468–473, 1994.

[171] Oliver Kramer. A Review of Constraint-Handling Techniques for Evolution Strategies. *Appl. Comput. Intell. Soft Comput.*, 2010(1):1–11, 2010.

[172] Walker H. Land, Xingye Qiao, Dan Margolis, and Ron Gottlieb. A new tool for survival analysis: evolutionary programming/evolutionary strategies (EP/ES) support vector regression hybrid using both censored / non-censored (event) data. *Procedia Comput. Sci.*, 6:267–272, jan 2011.

[173] W B Langdon. Size Fair and Homologous Tree Genetic Programming Crossovers. *Proc. 1st Genet. Evol. Comput. Conf. GECCO 1999*, 2:1092–1097, 1999.

[174] W. B. Langdon and B. F. Buxton. Genetic programming for mining DNA chip data from cancer patients. *Genet. Program. Evolvable*, 5(3):251–257, 2004.

[175] W B Langdon and J P Nordin. Seeding genetic programming populations. *Genet. Program. Proc.*, 1802:304–315, 2000.

[176] W B Langdon and R Poli. Why Ants are Hard. *Genet. Program. 1998 Proc. Third Annu. Conf.*, (CSRP-98-4):193–201, 1998.

[177] W. B. Langdon and Riccardo Poli. *Foundations of Genetic Programming*. Springer-Verlag.

[178] WB Langdon. *Genetic Programming and Data Structures*. PhD thesis, 1996.

[179] W.B Langdon. Scheduling maintenance of electrical power transmission networks using genetic programming. *Iee Power Ser.*, 1997.

[180] W.B. Langdon. *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!* Springer US, 1998.

[181] Wb Langdon and R Poli. Better Trained Ants for Genetic Programming. 44:1–9, 1998.

[182] Laupacis A, Wells G, Richardson WS, Tugwell P. Guides to the Medical Literature: V. How to Use an Article About Prognosis. *JAMA*, 272(3):234–237, 1994.

[183] S Lavington, N Dewhurst, E Wilkins, and A Freitas. Interfacing knowledge discovery algorithms to large database management systems. 41:605–617, 1999.

[184] CT Le and BL Lindgren. Computational Regression Implementation of the Conditional Logistic Model in the Analysis of Epidemiologic Matched Studies. *Comput Biomed Res*, 21(1):48–52, 1988.

[185] Elisa T Lee, Barbara V Howard, Wenyu Wang, Thomas K Welty, James M Galloway, Lyle G Best, Richard R Fabsitz, Ying Zhang, Jeunliang Yeh, and Richard B Devereux. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. *Circulation*, 113(25):2897–905, jun 2006.

[186] T.L. Lew, a.B. Spencer, F. Scarpa, K. Worden, a. Rutherford, and F. Hemez. Identification of response surface models using genetic programming. *Mech. Syst. Signal Process.*, 20(8):1819–1831, nov 2006.

[187] Li Li, Wei Jiang, Xia Li, Kathy L Moser, Zheng Guo, Lei Du, Qiuju Wang, Eric J Topol, Qing Wang, and Shaoqi Rao. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85(1):16–23, jan 2005.

[188] Linxia Liao and Radu Pavel. Machinery time to failure prediction - Case study and lesson learned for a spindle bearing application. *PHM 2013 - 2013 IEEE Int. Conf. Progn. Heal. Manag. Conf. Proc.*, 2013.

[189] S M Liew, J Doust, and P Glasziou. Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart*, 97(9):689–97, may 2011.

[190] Ricardo Linden and Amit Bhaya. Evolving fuzzy rules to model gene expression. *Biosystems.*, 88(1-2):76–91, mar 2007.

[191] MA Michael A Lones. Enzyme Genetic Programming: Modelling Biological Evolvability in Genetic Programming. *Development*, (September), 2003.

[192] S Ludwig and S Roos. Prognosis of breast cancer using genetic programming. In Rossitza Setchi, Ivan Jordanov, RobertJ. Howlett, and LakhmiC. Jain, editors, *Knowledge-Based Intell. Inf. Eng. Syst.*, volume 6279 of *Lecture Notes in Computer Science*, pages 536–545. Springer Berlin Heidelberg, 2010.

[193] Simone a. Ludwig. Prediction of breast cancer biopsy outcomes using a distributed genetic programming approach. *Proc. ACM Int. Conf. Heal. informatics - IHI '10*, page 694, 2010.

[194] Sean Luke. *Essentials of Metaheuristics*. Lulu, 2009.

[195] Sean Luke. *Essentials of Metaheuristics*. Lulu, second edition, 2013.

[196] Sean Luke and Liviu Panait. Lexicographic Parsimony Pressure. *GECCO 2002 Proc. Genet. Evol. Comput. Conf.*, pages 829–836, 2002.

[197] Sean Luke and Lee Spector. A Revised Comparison of Crossover and Mutation in Genetic Programming. *Genet. Program. 1998 Proc. Third Annu. Conf.*, pages 208–213, 1998.

[198] Robert M MacCallum. Introducing a Perl Genetic Programming System: and Can Meta-evolution Solve the Bloat Problem? *EuroGP*, 2610:364–373, 2003.

[199] AG Mainous, R Baker, RJ Koopman, A Saxena, VA Diaz, CJ Everett, and A Majeed. Impact of the population at risk of diabetes on projections of diabetes burden in the United States: an epidemic on the way. *Diabetologia*, 50(5):934–40, may 2007.

[200] Arch G Mainous, Richelle J Koopman, Vanessa a Diaz, Charles J Everett, Peter W F Wilson, and Barbara C Tilley. A coronary heart disease risk score based on patient-reported information. *Am. J. Cardiol.*, 99(9):1236–41, may 2007.

[201] S. R. Maxwell. Why Might Some Problems Be Difficult for Genetic Programming to Find Solutions? In *Late Break. Pap. Genet. Program. 1996 Conf. Stanford Univ. July 28-31, 1996*, pages 125–128, 1996.

[202] Aoife C McGovern, David Broadhurst, Janet Taylor, Naheed Kaderbhai, Michael K Winson, David a Small, Jem J Rowland, Douglas B Kell, and Royston Goodacre. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. *Biotechnol. Bioeng.*, 78(5):527–38, jun 2002.

[203] Ben McKay, Mark J. Willis, and Geoffrey W. Barton. Using a Tree Structured Genetic Algorithm to Perform Symbolic Regression. In *First Int. Conf. Genet. Algorithms Eng. Syst. Innov. Appl. GALESIA*, pages 487–492, 1995.

[204] Nicholas Freitag McPhee and Riccardo Poli. A schema theory analysis of the evolution of size in genetic programming with linear representations. In *Genet. Program. Proc. EuroGP'2001*, volume 2038, pages 108–125, 2001.

[205] Olaf Mersmann. emoa: Evolutionary Multiobjective Optimization Algorithms, 2012.

[206] Z. Michalewicz and M. Schoenauer. Evolutionary algorithms for constrained parameter optimization problems. *Evol. Comput.*, 4(1):1–32, 1996.

[207] Zbigniew Michalewicz. A Survey of Constraint Handling Techniques in Evolutionary Computation Methods. In J. R. McDonnell, R. G. Reynolds, and D. B. Fogel, editors, *Evol. Program. IVProceedings Fourth Annu. Conf. Evol. Program.*, pages 135–155. MIT Press, 1 edition, 1995.

[208] Zbigniew Michalewicz. Genetic Algorithms, Numerical Optimization, and Constraints. *Proc. sixth Int. Conf. Genet. algorithms*, 195:151–158, 1995.

[209] Zbigniew Michalewicz and Girish Nazhiyath. Genocop III: A Co-evolutionary Algorithm for Numerical Optimization Problems with Nonlinear Constraints. *Proc. Second IEEE Int. Conf. Evol. Comput.*, pages 647–651, 1995.

[210] David J. Montana. Strongly Typed Genetic Programming. *Evol. Comput.*, 3(2):199–230, 1995.

[211] Jason H. Moore, Joel S. Parker, Nancy J. Olsen, and Thomas M. Aune. Symbolic discriminant analysis of microarray data in autoimmune disease. *Genet. Epidemiol.*, 23(1):57–69, 2002.

[212] JH Moore, JS Parker, and LW Hahn. Symbolic discriminant analysis for mining gene expression patterns. *Mach. Learn. ECML 2001. Lect. Notes Comput. Sci.*, 2167:372–381, 2001.

[213] Alison a Motsinger, Stephen L Lee, George Mellick, and Marylyn D Ritchie. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics*, 7:39, jan 2006.

[214] KM Narayan, EW Gregg, A Fagot-Campagna, MM Engelgau, and F Vinicor. Diabetes–a common, growing, serious, costly, and potentially preventable public health problem. *Diabetes Res Clin Pr.*, 50(Suppl 2):S77–84, oct 2000.

[215] S a M Nashef, F Roques, P Michel, E Gauducheau, S Lemeshow, and R Salamon. European system for cardiac operative risk evaluation (Euro SCORE) q. 16:9–13, 2010.

[216] National Institute for Health and Clinical Excellence. Lipid Modification : Cardiovascular risk assessment and the modification of blood lipids for the primary

and secondary prevention of cardiovascular disease. *London NICE*, (May), 2008.

[217] National Institute for Health and Clinical Excellence. Lipid Modification: Cardiovascular risk assessment and the modifications of blood lipids for the primary and secondary prevention of cardiovascular disease. Guideline 67. *London Natl. Inst. Heal. Clin. Excell.*, 2008(March), 2010.

[218] National Vascular Disease Prevention Alliance. Guidelines for the Assessment of Absolute Cardiovascular Disease Risk. *Melb. Natl. Stroke Found.*, 2012.

[219] NHS Choices. About the National Health Service (NHS), 2012.

[220] NHS Connecting for Health. OPCS-4 Classification, 2012.

[221] NHS Connecting for Health. Read Codes, 2012.

[222] NHS Information Centre for Health and Social Care. General Practice Trends in the UK. (March):1–11, 2012.

[223] NHS Information Centre for Health and Social Care. Health Episode Statistics Online. nov 2012.

[224] NHS Primary Care Data Quality Team. READ CODES: A GUIDE FOR GENERAL PRACTICE. Technical report, 2010.

[225] Orazio Nicolotti, Valerie J. Gillet, Peter J. Fleming, and Darren V S Green. Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. *J. Med. Chem.*, 45(23):5069–5080, 2002.

[226] Office for National Statistics. Mortality statistics : Metadata. *Off. Natl. Stat.*, (October), 2012.

[227] Steve O'Hagan, Warwick B Dunn, Marie Brown, Joshua D Knowles, and Douglas B Kell. Closed-Loop , Multiobjective Optimization of Analytical Instrumentation : Gas Chromatography / Time-of-Flight Mass Spectrometry of

the Metabolomes of Human Serum and of Yeast Fermentations. *Anal. Chem*, 77(1):290–303, 2005.

[228] A Ohinmaa, P Jacobs, S Simpson, and J Johnson. The projection of prevalence and cost of diabetes in Canada: 2000 to 2016. *Can J Diabetes.*, 28(2):1–8, 2004.

[229] L Ohno-Machado. A COMPARISON OF COX PROPORTIONAL HAZARDS AND ARTIFICIAL NEURAL NETWORK MODELS FOR MEDICAL PROGNOSIS. *Comput. Biol. Med.*, 27(1):55–65, 1997.

[230] Michael O'Neill and Conor Ryan. *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*.

[231] Michael O'Neill, Leonardo Vanneschi, Steven Gustafson, and Wolfgang Banzhaf. Open issues in Genetic Programming. *Genet. Program. Evolvable Mach.*, 11(3-4):339–363, 2010.

[232] Oxford Center for Evidence-Based Medicine. Glossary of terms in Evidence-Based Medicine.

[233] Luigi PALMIERI, Salvatore PANICO, Diego VANUZZO, Marco FERRARIO, Lorenza PILOTTO, Roberto SEGA, Giancarlo CESANA, e Simona GIAMPAOLI, and per il Gruppo di Ricerca del Progetto CUORE. La valutazione del rischio cardiovascolare globale assoluto: il punteggio individuale del progetto CUORE. *Ann Ist Super Sanità*, 40(4):393–399, 2004.

[234] D Parrott, X Li, and V Ciesielski. Multi-objective Techniques in Genetic Programming for Evolving Classifiers. *Proc. 7th IEEE Congr. Evol. Comput. CEC 2005*, 2:1141–1148, 2005.

[235] Prabasaj Paul, Michael L. Pennell, and Stanley Lemeshow. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat. Med.*, 32(1):67–80, 2013.

[236] P Peduzzi and J Concato. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48(12):1503–1510, 1995.

[237] E Pellegrini, M Maurantonio, I M Giannico, M S Simonini, D Ganazzi, L Carulli, R D'Amico, A Baldini, P Loria, M Bertolotti, and N Carulli. Risk for cardiovascular events in an Italian population of patients with type 2 diabetes. *Nutr Metab Cardiovasc Dis*, 21(11):885–92, nov 2011.

[238] M. Podbregar, M. Kovačič, A. Podbregar-Marš, and M. Brezocnik. Predicting defibrillation success by 'genetic' programming in patients with out-of-hospital cardiac arrest. *Resuscitation*, 57(2):153–159, may 2003.

[239] R. Poli, W. B. Langdon, and N. F. McPhee. *A field guide to genetic programming*. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008.

[240] Riccardo Poli and Mario Graff. There is a free lunch for hyper-Heuristics, genetic programming and computer scientists. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 5481 LNCS:195–207, 2009.

[241] Riccardo Poli, Mario Graff, McPhee, and Nicholas Freitag. Free Lunches for Function and Program Induction. In *Proc. Tenth ACM SIGEVO Work. Found. Genet. Algorithms*, pages 183–194, 2009.

[242] Riccardo Poli and W B Langdon. A New Schema Theorem for Genetic Programming with One-point Crossover and Point Mutation 2 Previous Work on Schemata for GP. pages 1–25, 1997.

[243] Riccardo Poli and W B Langdon. On the Search Properties of Different Crossover Operators in Genetic Programming. *Proc. Genet. Program.*, pages 147–155, 1998.

[244] Riccardo Poli and William B. Langdon. Schema Theory for GP with One-point Crossover and Point Mutation. *Evol. Comput.*, 6(3):231–252, 1998.

[245] The Stroke Association Prepared by: British Cardiac Society, British Hypertension Society, Diabetes UK, HEART UK, Primary Care Cardiovascular Society. JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart*, 91(Suppl 5):v1–v52, dec 2005.

[246] QResearch. QResearch, 2014.

[247] R Core Development Team. R: A Language and Environment for Statistical Computing, 2013.

[248] P M Ravdin, G M Clark, S G Hilsenbeck, M a Owens, P Vendely, M R Pandian, and W L McGuire. A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Res. Treat.*, 21(1):47–53, jan 1992.

[249] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Phd thesis, Technical University of Berlin, 1971.

[250] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, 1973.

[251] David M Reif, Bill C White, and Jason H Moore. No Title. *Integr. Anal. Genet. genomic proteomic data*, 1(1):67–75, 2004.

[252] Jon T. Richardson, Mark R. Palmer, Gunar E. Liepins, and Mike R. Hilliard. Some Guidelines for Genetic Algorithms with Penalty Functions. In *Proc. 3rd Int. Conf. Genet. Algorithms*, pages 191–197. Morgan Kaufmann Publishers Inc, 1989.

[253] Paul M Ridker, Julie E Buring, Nader Rifai, and Nancy R Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA*, 297(6):611–9, feb 2007.

[254] Paul M Ridker, Nina P Paynter, Nader Rifai, J Michael Gaziano, and Nancy R Cook. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation*, 118(22):2243–51, 4p following 2251, nov 2008.

[255] B Ripley and R Ripley. Neural networks as statistical methods in survival analysis. In Richard Dybowski and Vanya Gant, editors, *Clin. Appl. Artif. Neural Networks*, pages 237–255. 2001.

[256] Marylyn D Ritchie, Bill C White, Joel S Parker, Lance W Hahn, and Jason H Moore. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 4:28, jul 2003.

[257] MD Ritchie, AA Motsinger, WS Bush, C S Coffey, and JH Moore. Genetic programming neural networks: A powerful bioinformatics tool for human genetics. *Appl Soft Comput.*, 7(1):471–479, 2007.

[258] Katya Rodríguez-vázquez, Carlos M Fonseca, and Peter J Fleming. Multiobjective genetic programming: a nonlinear system identification application. In *Late Break. Pap. 1997 Genet. Program. Conf.*, pages 207–212, 1997.

[259] Justinian Rosca. Generality Versus Size in Genetic Programming. In *ProceedingGenetic Program.*, pages 381–387, 1996.

[260] Justinian P Rosca. Analysis of Complexity Drift in Genetic Programming. In *Genet. Program. 1997 Proc. Second Annu. Conf.*, pages 286–294, 1997.

[261] W M Rosenberg and D L Sackett. On the need for evidence-based medicine. *J Public Heal. Med*, 17(3):330–4, 1995.

[262] P L Ross, P T Scardino, and M W Kattan. A catalog of prostate cancer nomograms. *J. Urol.*, 165(5):1562–8, may 2001.

[263] RJ Rubin, WM Altman, and DN Mendelson. Health care expenditures for people with diabetes mellitus, 1992. *J Clin Endocrinol Metab*, 78(4):809A–809F, 1994.

[264] Lars Ryden, Eberhard Standl, Malgorzata Bartnik, Greet Van den Berghe, John Betteridge, Menko-Jan de Boer, Francesco Cosentino, Bengt Jonsson, Markku Laakso, Klas Malmberg, Silvia Priori, Jan Ostergren, Jaakko Tuomilehto, Inga Thrainsdottir, Ilse Vanhorebeek, Marco Stramba-Badiale, Peter Lindgren, Qing Qiao, Silvia G Priori, Jean-Jacques Blanc, Andrzej Budaj, John Camm, Veronica Dean, Jaap Deckers, Kenneth Dickstein, John Lekakis, Keith McGregor, Marco Metra, Joao Morais, Ady Osterspey, Juan Tamargo, Jose Luis Zamorano, Jaap W Deckers, Michel Bertrand, Bernard Charbonnel, Erland Erdmann, Ele Ferrannini, Allan Flyvbjerg, Helmut Gohlke, Jose Ramon Gonzalez Juanatey, Ian Graham, Pedro Filipe Monteiro, Klaus Parhofer, Kalevi Pyorala, Itamar Raz, Guntram Schernthaner, Massimo Volpe, and David Wood. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). *Eur Hear. J*, 28(1):88–136, jan 2007.

[265] D J Sargent. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer*, 91(8 Suppl):1636–42, apr 2001.

[266] M. Schmidt and H. Lipson. Age-fitness pareto optimization. *Proc. Int. Conf. Genet. Evol. Comput.*, (2):543–544, 2010.

[267] F. Schmiedle, N. Drechsler, D. Grosse, and R. Drechsler. Priorities in multi-objective optimization for genetic programming. In *Proc. Genet. Evol. Comput. Conf.*, pages 1–170, 2001.

[268] Marc Schoenauer, Michele Sebag, Francois Jouve, Bertrand Lamy, and Habibou Maitournam. Evolutionary Identification of Macro-Mechanical Models. In *Adv. Genet. Program. 2*, volume 6, pages 467–488. 1996.

[269] G Schwarzer, Werner Vach, and Martin Schumacher. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat. Med.*, pages 541–561, 2000.

[270] Christian Setzkorn. On The Use Of Multi-Objective Evolutionary Algorithms For Classification Rule Induction. (March):253, 2005.

[271] Christian Setzkorn, Azzam F G Taktak, and Bertil E Damato. On the use of multi-objective evolutionary algorithms for survival analysis. *Biosystems.*, 87(1):31–48, jan 2007.

[272] Shital C Shah and Andrew Kusiak. Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.*, 31(3):183–96, jul 2004.

[273] a D Shaw, M K Winson, a M Woodward, a C McGovern, H M Davey, N Kaderbhai, D Broadhurst, R J Gilbert, J Taylor, E M Timmins, R Goodacre, D B Kell, B K Alsberg, and J J Rowland. Rapid analysis of high-dimensional bioprocesses using multivariate spectroscopies and advanced chemometrics. *Adv. Biochem. Eng. Biotechnol.*, 66:83–113, jan 2000.

[274] Elizabeth Shephard, Sally Stapley, and William Hamilton. The use of electronic databases in primary care research. *Fam. Pract.*, 28(4):352–4, aug 2011.

[275] Sara Silva and Ernesto Costa. Dynamic limits for bloat control in genetic programming and a review of past and current bloat theories. *Genet. Program. Evolvable Mach.*, 10(2):141–179, 2009.

[276] P C Simons, a Algra, M F van de Laak, D E Grobbee, and Y van der Graaf. Second manifestations of ARTerial disease (SMART) study: rationale and design. *Eur. J. Epidemiol.*, 15(9):773–81, oct 1999.

[277] Judith D. Singer and John B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, 2003.

[278] Moshe Sipper. *Evolved to Win*. Lulu, 2011.

[279] Alice E Smith and David W Coit. Constraint handling techniques: penalty functions. In T. Back, D.B. Fogel, and Z. Michalewicz, editors, *Handb. Evol. Comput.*, chapter C 5.2. Oxford University Press and Institute of Physics Publishing, Oxford, 1997.

[280] S. H. Song and P. M. Brown. Coronary heart disease risk assessment in diabetes mellitus: comparison of UKPDS risk engine with Framingham risk assessment function and its clinical implications. *Diabet Med*, 21(3):238–245, mar 2004.

[281] Lee Spector and Alan Robinson. Genetic Programming and Autoconstructive Evolution with the Push Programming Language. *Gpem*, 3(1):7–40, 2002.

[282] N Srinivas and Kalyanmoy Deb. Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms 1 Introduction. *J. Evol. Comput.*, 2(3):221–248, 1994.

[283] Jeffrey W Stephens, Gareth Ambler, Patrick Vallance, D John Betteridge, Steve E Humphries, and Steven J Hurel. Cardiovascular risk and diabetes. Are the methods of risk prediction satisfactory? *Eur J Cardiovasc Prev Rehabil*, 11(6):521–528, 2004.

[284] R J Stevens, V Kothari, a I Adler, and I M Stratton. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56); United Kingdom Prospective Diabetes Study (UKPDS) Group. *Clin Sci*, 101(6):671–9, dec 2001.

[285] E. W. Steyerberg, M. J. C. Eijkemans, F. E. Harrell, and J. D. F. Habbema. Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. *Med. Decis. Mak.*, 21(1):45–56, feb 2001.

[286] Ewout W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, 2009.

[287] a Taktak, L Antolini, M Aung, P Boracchi, I Campbell, B Damato, E Ifeachor, N Lama, P Lisboa, C Setzkorn, V Stalbovskaya, and E Biganzoli. Double-blind evaluation and benchmarking of survival models in a multi-centre study. *Comp. Biol. Med.*, 37(8):1108–20, aug 2007.

[288] RL Tannen, MG Weiner, and Dawei Xie. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ*, 338, 2009.

[289] A Rosemary Tate, Alexander G R Martin, Aishath Ali, and Jackie A Cassell. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. *BMJ Open*, 1(1):e000025, jan 2011.

[290] J Taylor, R Goodacre, WG Wade, JJ Rowland, and DB Kell. The deconvolution of pyrolysis mass spectra using genetic programming. *Fems Microbiol. Lett.*, 160(2):237–246, 1998.

[291] A Teller and M Veloso. PADO: A new learning architecture for object recognition. *Symb. Vis. Learn.*, 4:81–116, 1996.

[292] R Thadhani. Formal trials versus observational studies. In A Mehta, M Beck, and G Sunder-Plassmann, editors, *Fabry Dis. Perspect. from 5 Years FOS*, chapter 14. Oxford PharmaGenesis, 2006.

[293] The Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*, 375(9733):2215–2222, 2010.

[294] The Health Improvement Network. THIN, 2014.

[295] The National Collaborating Centre for Chronic Conditions. Type 2 Diabetes National Clinical Guideline for Management in Primary and Secondary Care (update). *London R. Coll. Physicians*, 2008.

[296] Krish Thiru, Alan Hassey, and Frank Sullivan. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ*, 326(7398):1070, 2003.

[297] Athanasios Tsakonas, Georgios Dounias, Jan Jantzen, Hubertus Axer, Beth Bjerregaard, and Diedrich Graf von Keyserlingk. Evolving rule-based systems in two medical domains using genetic programming. *Artif. Intell. Med.*, 32(3):195–216, nov 2004.

[298] UK Prospective Diabetes Study Group. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. *Diabetologia*, 34(12):877–890, 1991.

[299] UK Prospective Diabetes Study (UKPDS) Group. Itensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet*, 352(Ukpds 33):837–853, 1998.

[300] H Uno, T Cai, MJ Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat Med*, 30(10):1105–1117, 2011.

[301] Seetharaman Vaidyanathan, David I Broadhurst, Douglas B Kell, and Royston Goodacre. Explanatory optimization of protein mass spectrometry via genetic search. *Anal. Chem.*, 75(23):6679–86, dec 2003.

[302] Amber a W a van der Heijden, Monica M Ortegon, Louis W Niessen, Giel Nijpels, and Jacqueline M Dekker. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. *Diabetes Care*, 32(11):2094–8, nov 2009.

[303] S van Dieren, J W J Beulens, a P Kengne, L M Peelen, G E H M Rutten, M Woodward, Y T van der Schouw, and K G M Moons. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart*, 98(5):360–9, mar 2012.

[304] S van Dieren, L M Peelen, U Nöthlings, Y T van der Schouw, G E H M Rutten, a M W Spijkerman, D L van der A, D Sluik, H Boeing, K G M Moons, and J W J Beulens. External validation of the UK Prospective Diabetes Study (UKPDS) risk engine in patients with type 2 diabetes. *Diabetologia*, 54(2):264–70, feb 2011.

[305] JP Vandenbroucke. Prospective or retrospective: What's in a name ? *BMJ*, 302(6771):249–250, 1991.

[306] Leonardo Vanneschi, Antonella Farinaccio, Giancarlo Mauri, Mauro Antoniotti, Paolo Provero, and Mario Giacobini. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min.*, 4(1):12, jan 2011.

[307] E Vittinghoff. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer New York, 2005.

[308] C Henrik Westerberg and John Levine. Investigations of Different Seeding Strategies in a Genetic Planner. In *Appl. Evol. Comput.*, volume 2037, pages 505–514, 2001.

[309] P A Whigham. Grammatically-based genetic programming. *Proc. Work. Genet. Program. From Theory to Real-World Appl.*, 16(3):33–41, 1995.

[310] Malcolm D Whitfield, Michael Gillett, Michael Holmes, and Elaine Ogden. Predicting the impact of population level risk reduction in cardio-vascular disease and stroke on acute hospital admission rates over a 5 year period–a pilot study. *Public Health*, 120(12):1140–8, dec 2006.

[311] WHO. WHO factsheet No.37: CVD, 2011.

[312] King H. Wild S, Roglic G, Green A, Sicree R. Estimates for the year 2000 and projections for 2030. *Diabetes Care*, 27(5):1047–53, 2004.

[313] R Williams, L Van Gaal, and C Lucioni. Assessing the impact of complications on the costs of Type II diabetes. *Diabetologia*, 45(7):S13–7, jul 2002.

[314] P. W. F. Wilson, R. B. D'Agostino, D. Levy, a. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, 97(18):1837–1847, may 1998.

[315] D H Wolpert and W G Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1):67–82, 1997.

[316] AM Woodward, RJ Gilbert, and DB Kell. Genetic programming as an analytical tool for non-linear dielectric spectroscopy. *Bioelectrochemistry Bioenerg.*, 48(2):389–396, 1999.

[317] M Woodward, X Zhang, F Barzi, W Pan, H Ueshima, A Rodgers, and S MacMahon. The effects of diabetes on the risks of major cardiovascular diseases and death in the Asia-Pacific region. *Diabetes Care*, 26(2):360–366, 2003.

[318] Mark Woodward, Peter Brindle, and Hugh Tunstall-Pedoe. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*, 93(2):172–6, mar 2007.

[319] World Health Organisation. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). 2010.

[320] Yangfeng Wu, Xiaoqing Liu, Xian Li, Ying Li, Liancheng Zhao, Zuo Chen, Yihe Li, Xuxu Rao, Beifan Zhou, Robert Detrano, and Kiang Liu. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation*, 114(21):2217–25, nov 2006.

[321] Lidia Yamamoto and Christian Tschudin. Experiments on the automatic evolution of protocols using genetic programming. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 3854 LNCS(October):13–28, 2006.

[322] Xilin Yang, Wing-Yee So, Alice P S Kong, Chung-Shun Ho, Christopher W K Lam, Richard J Stevens, Ramon R Lyu, Donald D Yin, Clive S Cockram, Peter C Y Tong, Vivian Wong, and Juliana C N Chan. Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry. *Diabetes Care*, 30(1):65–70, jan 2007.

[323] Jianjun Yu, Jindan Yu, Arpit a. Almal, Saravana M. Dhanasekaran, Debashis Ghosh, William P. Worzel, and Arul M. Chinnaiyan. Feature Selection and Molecular Classification of Cancer Using Genetic Programming. *Neoplasia*, 9(4):292–IN3, apr 2007.

[324] Bt Zhang, P Ohm, and H Muhlenbein. Evolutionary induction of sparse neural trees. *Evol. Comput.*, 5(2):213–36, 1997.

[325] Byoung-Tak Zhang and Heinz Mühlenbein. Balancing Accuracy and Parsimony in Genetic Programming. *Evol. Comput.*, 3(1):17–38, 1995.

[326] Byoung Tak Zhang and HeinzM Uhlenbein. Evolving Optimal Neural Networks Using Genetic Algorithms with Occam's Razor. *Complex Syst.*, 7(3):199–220, 1993.

[327] Yongqiang ZHANG and Huashan CHEN. Predicting for MTBF Failure Data Series of Software Reliability by Genetic Programming Algorithm. In *Proc. Sixth Int. Conf. Intell. Syst. Des. Appl.*, 2006.

[328] Yongqiang Zhang and Jingjie Yin. Software Reliability Model by AGP. In *Ind. Technol. 2008. ICIT 2008. IEEE Int. Conf.*, pages 1–5, 2008.

[329] B Zupan, J Demsar, M W Kattan, J R Beck, and I Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artif. Intell. Med.*, 20(1):59–75, aug 2000.